HUMAN GENOME VARIATIONS AND EVOLUTION WITH A FOCUS ON THE ANALYSIS OF TRANSPOSABLE ELEMENTS

Musaddeque Ahmed, M.Sc.

Department of Biological Sciences

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies, Brock University
St. Catharines, Ontario

To dad,

who I hope is smiling on me from above.

Abstract

Genome sequence varies in numerous ways among individuals although the gross architecture is fixed for all humans. Retrotransposons create one of the most abundant structural variants in the human genome and are divided in many families, with certain members in some families, e.g., L1, Alu, SVA, and HERV-K, remaining active for transposition. Along with other types of genomic variants, retrotransponson-derived variants contribute to the whole spectrum of genome variants in humans. With the advancement of sequencing techniques, many human genomes are being sequenced at the individual level, fueling the comparative research on these variants among individuals. In this thesis, the evolution and functional impact of structural variations is examined primarily focusing on retrotransposons in the context of human evolution. The thesis comprises of three different studies on the topics that are presented in three data chapters. First, the recent evolution of all human specific AluYb members, representing the second most active subfamily of Alus, was tracked to identify their source/master copy using a novel approach. All human-specific AluYb elements from the reference genome were extracted, aligned with one another to construct clusters of similar copies and each cluster was analyzed to generate the evolutionary relationship between the members of the cluster. The approach resulted in identification of one major driver copy of all human specific Yb8 and the source copy of the Yb9 lineage. Three new subfamilies within the AluYb family – Yb8a1, Yb10 and Yb11 were also identified, with Yb11 being the youngest and most polymorphic. Second, an attempt to construct a relation between transposable elements (TEs) and tandem repeats (TRs) was made at a genome-wide scale for the first time. Upon sequence comparison, positional cross-checking and other

relevant analyses, it was observed that over 20% of all TRs are derived from TEs. This result established the first connection between these two types of repetitive elements, and extends our appreciation for the impact of TEs on genomes. Furthermore, only 6% of these TE-derived TRs follow the already postulated initiation and expansion mechanisms, suggesting that the others are likely to follow a yet-unidentified mechanism. Third, by taking a combination of multiple computational approaches involving all types of genetic variations published so far including transposable elements, the first whole genome sequence of the most recent common ancestor of all modern human populations that diverged into different populations around 125,000-100,000 years ago was constructed. The study shows that the current reference genome sequence is 8.89 million base pairs larger than our common ancestor's genome, contributed by a whole spectrum of genetic mechanisms. The use of this ancestral reference genome to facilitate the analysis of personal genomes was demonstrated using an example genome and more insightful recent evolutionary analyses involving the Neanderthal genome. The three data chapters presented in this thesis conclude that the tandem repeats and transposable elements are not two entirely distinctly isolated elements as over 20% TRs are actually derived from TEs. Certain subfamilies of TEs themselves are still evolving with the generation of newer subfamilies. The evolutionary analyses of all TEs along with other genomic variants helped to construct the genome sequence of the most recent common ancestor to all modern human populations which provides a better alternative to human reference genome and can be a useful resource for the study of personal genomics, population genetics, human and primate evolution.

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor Prof. Ping Liang for his endless support, both scientifically and financially. I could never ask for a better mentor than him. I would also like to thank my other supervisors in the committee, Drs. Fiona Hunter and Andrew Reynolds for their guidance and support.

My heartiest gratitude goes for my wife Nayeema without whom I wouldn't be able to be here where I am now; and my mom, sister and brother who kept inspiring me from thousand miles away. I would also like to thank my peers and friends from the lab who were always there for me when I needed, particularly Daniel, Scott and Dr. Changhui Huang. I thank Dr. Wen Li as well for sharing her expertise with me on a project.

Lastly, and most importantly, I would like to thank two most important people in my life - my dad who always wanted me to reach for higher but couldn't wait to see me fulfill his dream; and my 2 year old son Zafir, who relentlessly brightens up my day when I go home no matter how hard the day was.

I thank the Almighty for everything I have.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

1KGP, 1000 Genome Project
BLAT, BLAST-Like Alignment Tool
CAHP, Common Ancestor to all Human Populations
cDNA, coding DNA
CEU, Utah residents with Northern and Western European ancestry
CHB, Han Chinese from Beijing
CNV, Copy Number Variation
dbRIP, Database of Retrotransposon Insertion Polymorphisms (dbRIP)
ERV, Endogenous Retrovirus
FoSTeS, Forkhead Stalling and Template Switching
HERV, Human ERV
hsYb, human-specific Alu Yb
JPT, Japanese
Kya, thousand years ago
LINE, Long INterspersed Element
MEI, Mobile Element Insertion
ML, molecular clock
mlTR, multi-locus TR
Mya, million years ago
NAHR, Non-Allelic Homogenous Recombination
NCBI, National Center for Biotechnology Information
NGS, Next Generation Sequencing
NHR, Non-Homologous Recombination
ORF, Open Reading Frame
PCR, polymerase chain reaction
PEM, Pair-End Mapping
RD, Read Depth of coverage
SINE, Short INterspersed Element
slTR, single-locus TR
SNP, Single Nucleotide Polymorphism
SNV, Single Nucleotide Variation
SR, Split Read
SV, Structural Variation
TE, Transposable Element
TIP, Transposon Insertion Polymorphism
TR, Tandem Repeat
TRDB, Tandem Repeat Database
TSD, Target site duplications
UCSC, University of California at Santa Cruz
UTR, UnTranslated Region
VNTR, Variable Number of Tandem Repeat
YRI, Yoruban

Chapter 1 : Introduction

(A part of this section is reprinted from the review article: Ahmed M, Liang P: Study of modern human evolution via comparative analysis with Neanderthal genome. 2013. *Genomics Inform* 11(4):230-238.)

Structural variations are one of the key features in comparative studies among individual human genomes and make up the focal point of this thesis. This chapter provides background information relevant to the subsequent chapters. **Chapter two** presents a study that tracks the recent evolution of one of the youngest subfamilies of transposons, the major type of genomic repeat elements that frequently cause structural variations. **Chapter three** presents a study that investigates the relation between transposable elements (TEs) and the other type of repeat elements, Tandem Repeats (TRs). **Chapter four** presents a study involving all structural variations among individual human genomes identified so far and the assembly of a genome sequence that represents the most recent common ancestor of all modern humans. Presentation and application of this proposed ancestral genome sequence along with a useful bioinformatics tool is also described in this chapter. **Chapter five** contains overall discussions and general conclusions for the entire thesis.

1.1 Variations in human genome

The haploid human genome is a linear chain of about 3 billion base pairs of DNA molecules in a single cell divided into 23 chromosomes. Even though, in general, the overall architecture, e.g., the number of chromosomes and the order of genes in the chromosomes, of the human genome is fixed among individuals, the primary DNA sequence may vary. Genomic variation is a natural phenomenon and can be defined as relative differences in DNA sequence or arrangements of stretches of sequences among individual genomes. While the concept of genomic variation is commonly made of the variations among individuals, the most common pathogenic result of genomic variation,

cancer, actually stems from variations between the cells in the same or different tissues or organs of a single individual (Beroukhim *et al.,* 2010). In a broader sense, genomic variation is meaningful only when compared among more than one individual genome. From a technical point of view, the genomic variations are detected and characterized by comparing an individual genome sequence with the sequence of the human reference genome (Lander *et al.*, 2001). The reference genome is the complete sequence of the genome obtained from 12 individuals as the representative "whole genome sequence" for the modern humans.

Genomic variations between two genome sequences can be of various types. The variations can be broadly categorized by their size and subcategorized by their mechanism of formation or sequence characteristics. The size of these variations ranges from single base pair changes, such as Single Nucleotide Variances (SNVs), to small insertions and deletions (indels), and to larger variations as in Structural Variants (SVs). SNVs are variations in a single base, and when observed frequently in or between population(s) with frequencies above 1%, they are called Single Nucleotide Polymorphisms (SNPs). Indels are insertions or deletions of sequences of 1-100bp in size. Structural variations are normally defined as larger than 100bp and can encompass millions of base pairs. Structural variations can occur from insertions or deletions of large genomic entities like Transposable Elements, Tandem Repeats, segmental duplication or deletion, processed pseudogenization or copy number variations (CNVs). SVs also include variations that are neutral in terms of size differences, such as *sequence inversions* where a block of sequence is reversely oriented or *balanced chromosomal*

translocation where two chromosomes exchange sequence with no consequential loss or gain of genetic materials (Alkan *et al.*, 2011).

Structural variation is a major source of genomic variation among different individuals. It has been identified in several studies that the total number of base pairs that differ between two individuals due to structural variation significantly surpasses the total number of SNPs (Conrad *et al.*, 2010; Lander *et al.*, 2001; Mills *et al.*, 2011). Many studies have been conducted in the last nine years to identify genome-wide structural variations either in individual genomes or in a large group of people (Chiang *et al.*, 2009; Feuk *et al.*, 2006; Kidd *et al.*, 2008; Korbel *et al.*, 2007; Lee *et al.*, 2008; McKernan *et al.*, 2009; Medvedev *et al.*, 2009; Redon *et al.*, 2006; Yoon *et al.*, 2009). The advent of high-throughput sequencing techniques and concurrent formations of large genome analysis consortia have considerably advanced the detection of novel structural variations in recent years. The structural variation team of the 1000 Genome Project performed the most comprehensive analysis to date and identified a large number of structural variations by analyzing the genome sequences of approximately 1,000 individuals from around the world (Mills *et al.*, 2011). With the cost of sequencing decreasing rapidly in recent years, studies are now conducted with a high coverage of DNA sequences, and a multitude of techniques are applied for the detection of structural variants (discussed in section 1.3). The challenge to combine all technologies to curate SVs is that each detection platform differs from one another by sequence quality, boundary resolution, and sensitivity (Ionita-Laza *et al.*, 2009). The biggest challenge so far is to pinpoint the exact boundary of a SV, as it requires highly sensitive detection techniques such as Split Read method with higher coverage of sequencing. However, the 1000 Genome Project provided the

exact breakpoints of the largest number of SVs to date. In a study presented in this thesis, a novel approach involving chimpanzee genome sequence is described to accurately pinpoint the boundaries of SVs for which only a genomic range could be estimated as potential boundary positions based on data in the original studies (described in chapter 4).

SVs are often generated by following three major mechanisms – non-allelic homologous recombination (NAHR), non-homologous end-joining or recombination (NHEJ/NHR), forkhead stalling and template switching (FoSTeS), and retrotransposition. Different types of SVs are formed by different mechanisms. For example, duplications often result from NAHR or NHR or transposition; deletions often result from NAHR or NHR; and inversions generally result from only NAHR, while FoSTeS generally forms a complex pattern involving deletion, duplication and inversion simultaneously at the same locus.

NAHR is a recombination event between highly homologous non-allelic sequences during cell division. When two repeat sequences in the same chromosome align and cross-over during meiosis, they can cause insertion or duplication or inversion depending on their orientation in each allele. Duplication or deletion occurs when the repeats are in direct orientation (head-to-head) to each other, and inversion may occur when they are in opposite directions (head-to-tail) (Figure 1.1). When the homologous regions (e.g., repeats) that cross over during cell division are located in different chromosomes, translocation takes place. NHR, on the other hand, takes place via cell repairing mechanisms in cases of DNA double stranded breaks due to some external influences such as radiation or internal influences such as V(D)J recombination. The exact molecular mechanism of NHR is still somewhat questionable, but they are found to take

place more often in repeat regions in the genome such as transposable elements. This characteristic can be explained as these transposable elements often co-reside with genomic regions that are vulnerable to DNA double-stranded breaks (Korbel *et al.*, 2007). The third major mechanism for SVs, FoSTeS, is a DNA replication-based mechanism that results in highly complex structural rearrangements. SVs resulting from this mechanism are very complex and they are not yet characterized by any specific definitive flanking sequence pattern or any bias towards any other genomic events, thus are not computationally detectable unlike the other three mechanisms.

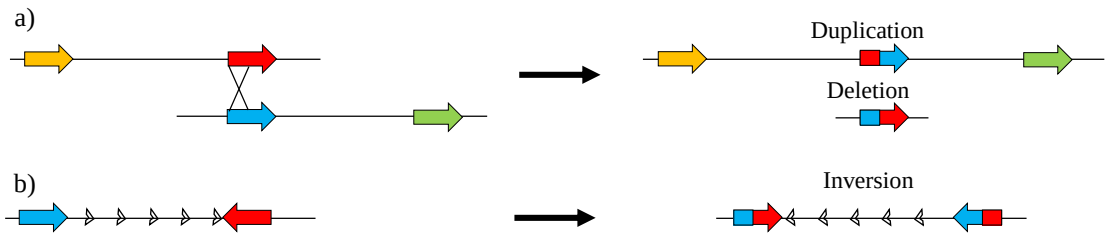


Figure 1.1 Genomic rearrangement by NAHR depends on the orientation of the homologous regions. Panel a illustrates how head-to-head orientation of the homologous regions may cause duplication or deletion, and panel b illustrates how head-to-tail orientation may cause inversion. NAHR, Non-Allelic Homologous Recombination.

1.2 Repeat elements, their evolution and contribution to genome variation

Repeats elements are the most abundant type of genomic sequences that can cause structural variations, making up a significant part of the entire human genome. Based on the relative positional relationship of the elements, they are divided into two broad categories –interspersed repeat elements or Transposable Elements (TEs) and tandemly

positioned DNA sequences or Tandem Repeats (TRs). Both types of elements are abundant in eukaryotes including humans (reviewed in Plohl *et al.*, 2012 and Rebollo *et al.*, 2012). They are discussed separately in the subsequent sections.

1.2.1 Transposable elements

Transposable genetics elements are DNA sequences that retain the ability to move from one genomic location to another. Since their identification in maize by Barbara McClintock (McClintock, 1956), they are found abundantly throughout many different genomes including plants and animals (Batzer *et al.*, 1993). TEs are very different from tandem repeats in many aspects, the foremost of which is that they are located sporadically throughout the genome as opposed to tandem repetition of sequences. TEs are often larger in size than tandem repeats. They are ubiquitous through mammalian genomes; as much as 45% and 37% of the entire human and mouse genome, respectively, is composed of TEs. Mammalian TEs can be broadly separated into two classes: DNA transposons and retrotransposons (Figure 1.2). DNA transposons are characterized by a cut-and-paste mechanism in which the sequence is directly transposed from one location to another by excision followed by insertion. The DNA transposon typically contains an open reading frame (ORF) within the sequence that encodes a transposase protein to facilitate the movement (Mizuuchi, 1992; van Luenen *et al.*, 1994). The transposition can be both autonomous and non-autonomous, and DNA transposons constitute about 3% of the entire human genome (Lander *et al.*, 2001), and they are no longer active in the human genome (reviewed in Feschette & McKinl, 2013). On the other hand, retrotransposons form a much larger group of TEs constituting approximately 42% of the entire human genome. The transposition mechanism of retrotransposons involves an

RNA intermediate to follow a copy-and-paste strategy as opposed to DNA transposons' cut-and-paste (Weiner *et al.*, 1986). The original copy of the TE (donor copy) is first transcribed to an RNA-intermediate, which is then reverse transcribed and inserted into a new genomic location to create a new copy (recipient copy) of the donor TE (reviewed in Levin & Moran, 2011). Among all TEs, retrotransposons are of particular interest because of their abundance and higher recent activity level than DNA transposons.

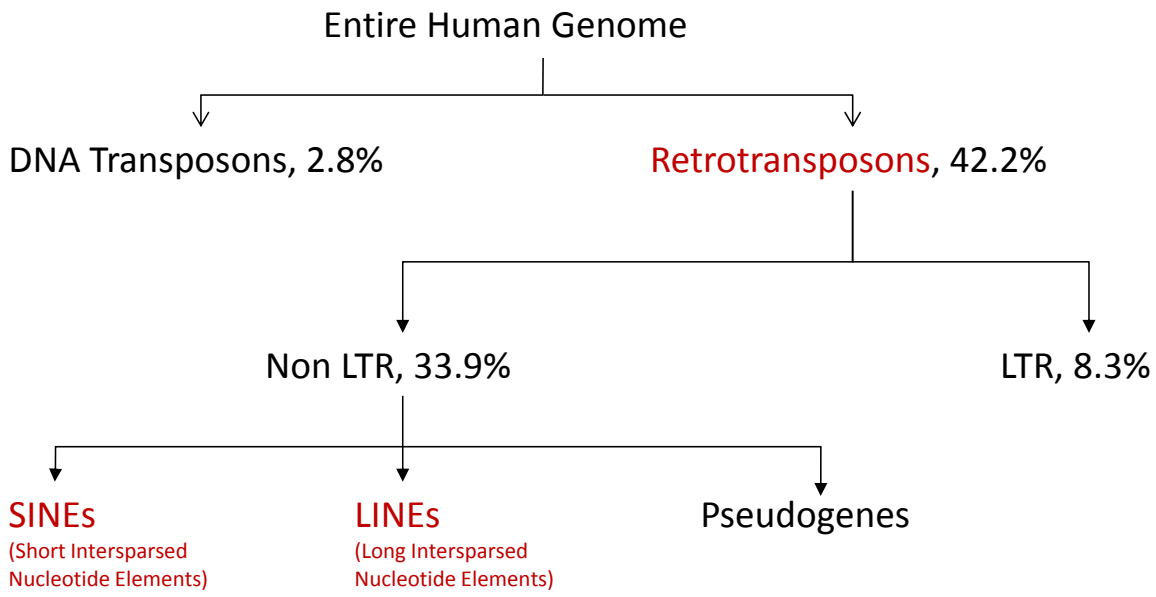


Figure 1.2 The categories and their hierarchy of transposable elements in the human genome.

There are three major types of retrotransposons in the human genome: retrovirus-like elements or Endogenous Retroviruses (ERVs), Short INterspersed Elements (SINEs) and Long INterspersed Elements (LINEs). ERVs are the longest among all retrotransposons ranging from 2-11 kbps in size, and they contain long terminal repeats and are capable of producing retrovirus-related protein much like retroviruses. However, ERVs lack the

envelope protein, which renders them incapable to travel between cells. LINEs are typically 4-6 kbps and form almost 21% of the entire human genome. LINEs are autonomous in the sense that they can translate their own mobilizing proteins (Boeke, 1997; Jurka, 1997; Mathias *et al.*, 1991). The major LINE family is L1 which is the only known active family of LINEs (Mills *et al.*, 2007). On contrary, SINEs are non-autonomous and rely on L1 machinery to transpose (Dewannieux *et al.*, 2003; Kajikawa & Okada, 2002). They are typically 300 bp in length and are the most successful type of retrotransposons with as many as 1 million copies constituting over 10% of the mass of the human genome (Batzer & Deininger, 2002; Lander *et al.*, 2001). Apart from these three major categories of retrotransposons, there is another relatively younger class of retrotransposons termed as SVA (SINE/VNTR/Alu). SVA elements only evolved about 25 million years ago in certain primate genomes and currently have ~3000 copies in the human genome (Wang *et al.*, 2005). SVA is a composite retrotransposon named after a SINE-R element, a variable number of tandem repeats (VNTR) and an Alu element (Shen *et al.*, 1994). A full-length SVA element is normally around 2 kb long and typically composed of a hexamer repeat region, the VNTR region, a HERV-K10 like region followed by poly-A tail (Ostertag *et al.*, 2003; Wang *et al.*, 2005). SVA elements are also non-autonomous, likely relying on L1 machinery to retro-transpose.

Even though there are many families of SINEs present in the human genome, Alu elements are the only transpositionally active kind (Mills *et al.*, 2007). Alu elements have shown tremendous proliferation throughout the evolution of primates and outnumber any other types of TEs. Moreover, Alu elements retain a higher activity level in recent times than any other TE making them more polymorphic for presence or absence across the

entire human population (Stewart *et al.*, 2011). These characteristics make Alu elements an intriguing subject to study and hence are one of the main foci in this thesis. Alu elements are dimeric nucleotide sequences with a left and right monomer joined by a poly-A region. Although there are over 1 million copies of Alu elements, comparatively only a small number of them are competent to mobilize – mostly from the youngest Alu family named AluY. These active progenitors of several other elements are termed "source" or "master" genes (Shen *et al.*, 1991). AluY family is an important retrotransposon family to study because of their current activity level, especially since new insertions are sometime related to diseases in humans. Approximately 11 Alu elements so far have been found to alter gene expression through exonization or exon skipping (Ferlini *et al.*, 1998; Ganguly *et al.*, 2003; Knebelmann *et al.*, 1995; Lev-Maor *et al.*, 2003; Ostertag *et al.*, 2003; Vervoort *et al.*, 1998). In at least two instances, an Alu inserted in the intron become exonized (retained) in the mature mRNA (Kreahling & Graveley, 2004; Lev-Maor *et al.*, 2003). The retention of exonized Alu element in the mRNA can lead to subtle differences in gene expression within individuals and/or populations. Furthermore, at least 14 Alu elements have been detected in exon, disrupting genes by causing reading frame shifts (Claverie-Martin *et al.*, 2003; Ostertag & Kazazian, 2001; Sukarova *et al.*, 2001). The study of the evolutionary trend of young Alu subfamilies and identifying still active Alu subfamilies can be particularly valuable in understanding the process of Alu expansion and also in population genetics due to their homoplasy-free nature. Chapter two describes a study where the evolution of one of the most active AluY subfamily, AluYb, was investigated and three novel subfamilies were proposed.

1.2.2 Tandem repeats

Initially termed as "junk DNA" together with other repetitive sequences, tandem repeats have gained some attention with the realization that their organization may provide some unique functional characteristics. Tandem repeats are organized by tandemly positioned homologous DNA sequences called *Repeat Units* in a head-to-tail pattern, which gives them their characteristic organization shared by many genomes. The centromeres of a vast array of species play a critical role in heterochromatin formation and chromosome segregation, and these centromeres are highly enriched with TRs (Morris & Moazed, 2007). Many of the functions demonstrated by TRs involve RNA interference mediated chromatin modification, which is important for heterochromatin formation (Alleman *et al.*, 2006; Chan *et al.*, 2006; Martienssen, 2003; Stam *et al.*, 2002).

Tandem repeats in human have widely varying repeat sizes, ranging from microsatellites of few base pairs to megasatellites which can be several hundred base pairs in size (Ames *et al.*, 2008; Gelfand *et al.*, 2007). Microsatellites now have extensive use as genetic markers in forensics (Hagelberg *et al.*, 1991; Olaisen *et al.*, 1997) and genomic mapping (Dib *et al.*, 1996; Dietrich *et al.*, 1996, Armour *et al.*, 1996; Bowcock *et al.*, 1994). Furthermore, expansion of microsatellites has been associated with many genetic diseases and the level of variability in VNTR can be an indicator of predisposition of several forms of diseases including cancer (Mandel, 1997; Wada *et al.*, 1994). Despite the critical function of microsatellites (briefly discussed in section 3.1), little is known about their origin and mechanism of formation.

Even though the two distinct types of repeat elements, tandem repeats and transposable elements, are ubiquitous and have been extensively studied, little has been

done to seek any association between the two. Chapter three of this thesis presents a study that established a connection between the two types of repeats to the sequence level for the first time.

1.3 Identification of genomic variants and their use in modern human evolutionary study

One of the biggest questions in evolutionary biology is what makes the modern humans modern. With advancements in all sectors starting from paleontology to molecular biology, the evolution of humans is much better understood now than ever before. Since the divergence of humans from its closest extant lineage, chimpanzee, six million years ago (Goodman *et al.*, 1998), the human genome has evolved independently accumulating its own unique changes. The archaic humans were a lot different than the modern humans, even the humans ~200,000 years ago are phenotypically distinctive from modern-day humans (Pearce *et al.*, 2013). Many of these phenotypic changes are brought about as a result of adaptive genomic variations. A study estimates that the proportion of genomic differences between humans and chimps to be 6.59%, with 5.07% differences due to indels (Wetterbom *et al.*, 2006). This denotes the importance of detecting genomic structural variations, either small or large, in evolutionary studies as well as comparative population genetics to seek for indel-mediated etiology of our phenotypic differences. Either between humans and chimp or between different populations, differences in coding regions are minimal although the genes with redundant functionality, such as binding proteins or finger motifs, are duplicated many times (Korbel *et al.*, 2007). Thus the differences observed in non-coding regions are likely the

major evolutionary force. Though there are many mechanisms for creating indels in non-coding regions, transposable elements are considered the major driving force to cause the differences between two species or two populations of the species. With almost half of the entire human genome made up of transposable elements (http://genome.ucsc.edu), insertions of new TEs or TE-mediated genomic variations can cause significant change in genomic plasticity subsequently causing phenotypic anomalies. Thus it is extremely important to study the evolutionary expansion pattern of transposable elements, what triggers their activity level or what transposable elements are specific to a species or a particular group within a species. The study described in Chapter four proposed the genome sequence of the most recent common ancestor of all modern human populations taking all sequence variations identified so far in consideration. The study provides a valuable resource for all future studies involving individual genome sequencing, as well as a comprehensive picture of expansion of the major families of retrotransposons in recent times of human evolution.

One of the major driving factors that enables the construction of the genome sequence of the most recent common ancestors of all modern humans is the availability of structural variations data between individuals from all major human populations. Thus, the advent in Next Generation Sequencing (NGS) and identification techniques of these variations plays a big role towards obtaining these data. Even eight to ten years ago, the major methods to identify structural variation were microarray-based and comparative genomic hybridization. With the establishment of the next generation sequencing technologies, it soon became the most effective way of detecting SVs in large scale. Tuzun et al. (2005) was among the first research groups to utilize paired-end sequences to

detect SVs using bioinformatics methods, but the resolution and accuracy was much lower than today's sequence-based SV detection quality (Tuzun *et al.*, 2005). Next generation sequencing refers to the current sequencing techniques and instruments, which is in broader sense actually the second generation sequencing, with Sanger's capillary DNA sequencing being the first generation. With the cost of sequencing coming down to $0.5 per million base pairs, next generation sequencing is the future of genome analysis.

The fundamental idea behind next generation sequencing is that the whole genome is broken down in to small fragments and each fragment is sequenced in smaller stretches multiple times to create overlapping sequences (Metzker, 2010). The output of most second generation sequencing techniques is comprised of hundreds of millions of DNA sequences of size ranging from 50 to several hundred base pair DNA snippets (*reads*). The average number of reads that overlap each position of the whole genome is termed as *Sequence Coverage*, for example, a genome sequenced with 30x coverage means that each base pair of the genome is present in on average 30 reads. The higher the coverage is, the better the sensitivity for downstream analysis, and with current technology, 40x coverage is very accessible which is good enough to conduct highly sensitive comparative analyses. Depending on read length, sequencing platform and coverage, there are three established methods for discovering SVs – paired end mapping (PEM), Read depth of coverage (RD) and split-read method (SR). All of these methods are applied after the reads are mapped against a reference genome.

Paired End Mapping is applicable when a pair of sequence reads come from two ends of a single DNA fragment of experimentally selected size, normally around 300bp for PCR-based NGS platforms, such as Illumina. The read length depends on the sequencing

platform. The basis of PEM technique is that a DNA fragment spanning a junction point of a SV will have discordant mapping, i.e., the distance and/or orientation between the paired reads will be different than expected. For example, for a pair of reads of 35 bp each from a 300bp fragment, the gap between the pair mates should be 230bp. In case of a deletion in the test genome compared to the reference genome, the reads will map further away than 230bp in the reference, or closer than 230bp in case of an insertion in the test genome.

Read depth of coverage technique is based on statistical analysis of density of reads per selected size of genomic area (termed *window*). In the method, the number of reads per window is calculated genome-wide, and any window that differs significantly than the genome average in terms of read density is likely to harbor a SV. RD technique is relatively more useful in detecting copy number aberrations than other computational approaches as any window containing a novel duplication will have higher read density than the other genome and vice versa for regions containing a deletion.

Split read technique makes use of the reads that physically span the junction point of a SV. If the test genome has an insertion or deletion compared to the reference genome, any read that spans across the junction will be split when mapped to the reference genome and the two pieces are likely to map to different location. SR techniques are best suited for longer reads and high coverage sequencing, as the split sequence pieces can be long enough to map accurately to different loci in the reference genome.

While each of the three major detection techniques has its own advantage and disadvantage, PEM has been the widely used technique primarily because of short read

sizes at the early years of NGS era. SR provides the maximum resolution as it can call a SV to the exact nucleotide level. This can provide accurate information about the junction sequence, and if the SV is in a gene, the exact part of gene deleted or duplicated can be identified.

All of these computational approaches to detect sequence variations use a reference genome to align the reads and compare sequence with. The ancestral genome sequence described in Chapter four provides a better alternative than the currently used reference genome for reasons detailed in the related later section. Chapter four further describes the data presentation of the newly proposed genome sequence, as well as the use of this genome sequence in comparative bioinformatics by applying a combinatorial approach of PEM and SR on an individual's genome sequence. This novel genome sequence is also applied in evolutionary studies involving Neanderthal genome sequence to preliminary assess the progression of TE expansion in the current human genomes compared to the Neanderthals. The Neanderthal genome sequencing and its use in TE insertion polymorphism is discussed in the next subsection.

1.4 Using Neanderthal genome for studying recent evolution of TEs

(This subsection is reprinted from the review article: Ahmed M, Liang P: Study of modern human evolution via comparative analysis with Neanderthal genome. 2013. Genomics Inform 11(4):230-238.)

Modern humans are indeed a very young species compared to their cousins, evolving just about 200,000 years ago (ya), which is a fraction of the 6 million years since the divergence of the human and chimpanzee lineages (McDougall *et al.*, 2005). Fossils

suggest that modern humans first emerged in East Africa, and then fairly quickly spread all over the world in the next 185,000 years or so (Mellars, 2006). After the divergence of humans and chimps 6 million years ago (mya), the major landmark in human history is the emergence of bipedals about 4 mya, which enabled them to use their two front feet as hands. Many species evolved afterwards until the evolution of *Homo erectus*, who for the first time migrated out of Southern Africa and initiated the spread of humans all around the globe. The migrated population of *Homo erectus* in East Africa eventually gave rise to modern humans about 200,000 ya and to *Homo neanderthalensis* or Neanderthals about 400,000 ya (Hublin, 2009; Stringer & Hublin, 1999). Neanderthals survived until 28,000 ya, while modern humans are still surviving (Finlayson *et al.*, 2006). During the later part in their existence timespan, Neanderthals lived in Europe as well as in Western Asia and Middle East (Grun *et al.*, 2005; Krause *et al.*, 2007). Various lines of evidence suggest that modern humans started to migrate from East Africa to Europe and other parts of the world 100,000 ya, and the fossil evidence of humans and Neanderthals indicated that these species might have come into contact as early as 80,000 ya and co-habited for up to 10,000 years at certain geographic locations (Grun *et al.*, 2005).

In the field of human evolutionary biology, one of the most sought after questions has been what made modern humans superior and outcompete the other related species, i.e., the genomic features that are unique to humans. The whole genome sequencing of chimps, rhesus macaque and other primates has given considerable boosts in this field as the sequences of these primates opened up the possibility to conduct comprehensive comparative studies to the single nucleotide level (Chimpanzee Sequencing and Analysis Consortium, 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium *et*

al., 2007). Many attempts have been taken to identify the genetic reasons why modern humans developed such complex biological features than other primates, including the larger brain to body ratio, bipedalism, morphological changes and significant development of communication skill and cognitive behavior. Recent studies have used various statistical methods to compare the sequence of these primates with humans in an attempt to find human-specific genes and gene regulatory sequences eventually showing unexpectedly rapid evolution in the human lineage after the divergence from the ancestral primates (Bird *et al.*, 2007; Clark *et al.*, 2003; Haygood *et al.*, 2007; Pollard *et al.*, 2006; Prabhakar *et al.*, 2006; Prabhakar *et al.*, 2008). The results from these analyses exhibit a good overview of the human-specific genomic elements, but these results are unable to distinguish which of these human-specific elements are specific to modern humans only. Since there has been no complete genome sequence of any archaic humans until recently, such sequence comparisons are made only between modern human genome and other primates bypassing archaic humans, resulting in overwhelming number of differences and inability to identify which sequences changes are unique to modern humans and which are shared by all *Homo* species. Therefore, the comparative analysis between modern humans and archaic humans is expected to be more interesting and valuable by being more effective in identifying the critical genes and/or regulatory elements that may be fully or partially responsible for the evolution of the modern humans over other humans.

Among all the Alu elements found in the entire human genome, only about 0.5% are found to be present in human genome but absent in orthologous regions of other primates, thus identified as human-specific. This ‘young’ group of Alus is composed of

only about 5000 Alu elements that are believed to be integrated in the human genome after the divergence of humans and great apes (Batzer & Deininger, 1991; Batzer *et al.*, 1995a; Matera *et al.*, 1990; Roy *et al.*, 1999; Roy-Engel *et al.*, 2001). Studying the retrotransposon insertion loci in Neanderthals will suffice the identification of truly modern human-specific retrotransposon insertions. A similar comparative analysis would reveal other transposable elements, such as L1, SVAs or HERVs, that are specific to modern humans only, as well as those that are specific to Neanderthals. Retroelements are particularly important in population genetics. It is extremely rare that a newly inserted transposable element is completely excised, thus they act as a genetic fossil that are homoplasy-free. This identical-by-descent nature of retroelements makes them a better mean for population and evolutionary studies from SNPs in the sense that SNPs can be, though rarely, mutated back to the previous state. SNPs are also very hard to detect while handling ancient genome due to transformation and deamination (Briggs *et al.*, 2007), while RIPs are mainly presence or absence of a stretch of nucleotides. Once a retrotransposon is inserted in a new location in an individual, it becomes the subject of genetic drift. Over a short period, it starts spreading into the population. Depending on when a retroelement has integrated at a certain loci, it will be shared by different species or if recently enough, by different populations of the same species. Thus, RIPs occurring before the divergence of chimps and humans are shared by humans and chimps, but those occurring after are only present in humans. RIPs that are even more recent are specific to certain human populations only (Wang *et al.*, 2006a; Wang *et al.*, 2006b). The detailed information about all polymorphic retroelements and their frequency in different populations is extensively catalogued in the dbRIP database (Wang *et al.*, 2006b). The

identical-by-descent and homoplasy-free nature of RIPs make them useful genetic markers in population and evolutionary genetics. The specificity of RIPs can play a significant role in answering the question of admixture of Neanderthals and modern humans. Finding RIPs that are shared between Neanderthals and non-African populations, but not present in African population can be considered as a solid support for the proposed admixture between Neanderthals and non-African populations. In an ongoing study in our laboratory, over 500 RIPs are identified to be present in Khoisan and Bantu individuals who represent the oldest lineage of modern humans from Southern Africa but not in the reference human genome (unpublished). These oldest African lineage-specific RIPs theoretically should be absent from Neanderthals too based on the theory of admixture between Neanderthals and non-African human populations.

Chapter 2 : Identification of three new Alu Yb subfamilies by source tracking of recently integrated Alu Yb elements

(The content of this chapter is mostly copied from the published article: "Ahmed M, Li W, Liang P: Identification of three new Alu Yb subfamilies by source tracking of recently integrated Alu Yb elements. 2013. *Mobile DNA* 4:25" with the Materials and Method section moved before Result and Discussion section and some minor text edits.

The candidate is the main author of this article and was responsible for generating most of the data included in the article. The PCR amplification and gel purification described in subsection 2.4.3 was conducted by the second author while the primer design and post-sequencing analyses along with the other parts of the study were conducted by the candidate. The manuscript was drafted by the candidate and edited by the corresponding author, Dr. Liang, to its final form.)

2.1 Background

Alu elements are the most successful short interspersed elements (SINEs) in primate genomes. Alu elements have proliferated significantly throughout primate evolution and have expanded to more than 1 million copies in the human genome, constituting over 10% of the genome by mass (Batzer & Deininger, 2002; Lander *et al.*, 2001). The majority of these elements are suspected to have been inserted in the primate genome 35 to 60 million years ago, and since then the proliferation rate has reduced significantly by over 100 fold (Shen *et al.*, 1991). Thus, despite the large number of copies present in the human genome, only a small fraction of Alu elements are still active and capable of generating new copies (Mills *et al.*, 2006; Mills *et al.*, 2007; Wang *et al.*, 2006b). The activity of Alu elements has generated different subfamilies of varying ages, each subfamily being defined and characterized by a set of diagnostic mutations (Jurka & Milosavljevic, 1991). Each subfamily is thought to have expanded when its master or source copy accumulated a mutation and then actively transposed to new locations at different rates and time periods of evolution (Deininger *et al.*, 1992; Roy-Engel *et al.*, 2001).

The vast majority of the Alu elements currently found in the human genome were inserted before the divergence of humans and chimps, and thus are shared by all individuals of both species. The small fraction of Alu elements that have been recently inserted into the human genome are mostly restricted to several closely related young subfamilies, with the majority of these young elements being from the Ya5 and Yb8 Alu subfamilies (Batzer *et al.*, 1995b; Batzer *et al.*, 1996). Since almost all of these young

Alu elements were inserted into the human genome after the human–chimp divergence, they are only found in humans. Some of these subfamilies are so recent that they have members that are polymorphic for their presence or absence among individuals and/or populations (Arcot *et al.*, 1995; Carter *et al.*, 2004; Wang *et al.*, 2006a). The availability of a complete human reference genome and large quantities of individual genomic data from the 1000 Genome Project have facilitated the identification of these subfamilies and their level of polymorphism (Hormozdiari *et al.*, 2011; Stewart *et al.*, 2011). The homoplasy-free nature of Alu elements makes their polymorphic insertions very useful in phylogenetic studies, human population studies, forensics and DNA fingerprinting (Batzer & Deininger, 1991; Batzer *et al.*, 1994; Novick *et al.*, 1993; Novick *et al.*, 1995; Roy-Engel *et al.*, 2001).

Our study specifically focuses on human-specific Alu elements from the Yb lineage, mainly because they are the second largest young family by the number of copies in the human genome, comprising 40% of all human-specific Alu elements with more than 30% of these copies being polymorphic between individuals and/or populations, and also because they are amongst the most active TE subfamilies (Hedges *et al.*, 2004; Hormozdiari *et al.*, 2011; Stewart *et al.*, 2011). Alu Yb8 is the major subset of this family. Its human-specificity and high rate of being polymorphic among humans and its involvement in human diseases via de novo insertion suggest that this subfamily is still actively retrotransposing (Muratani *et al.*, 1991; Oldridge *et al.*, 1999). The Yb8 subfamily is characterized by a tandem duplication of seven nucleotides from the 246th to the 252nd position of the AluY consensus sequence. The concurrent mutation and transposition of certain Yb8 elements generated the Yb9 subfamily, which was the latest

Yb subfamily to be identified before this study and characterized by a C to G transversion at the 274th position (Roy-Engel *et al.*, 2001). In this study, using a computational approach we performed a genome-wide analysis of all human-specific Yb elements to identify their source copies and to track their recent evolutionary pathway. We successfully detected at least one driver copy for Yb8 and one Yb8 element that is potentially the source gene for the Yb9 subfamily. We also identified and characterized three new subfamilies in the Yb lineage: Yb8a1, Yb10, and Yb11. Yb11 is the youngest Yb subfamily reported to date.

2.2 Materials and Methods

2.2.1 Source copy tracking

All human-specific Yb elements were retrieved from a separate study (Tang *et al.*, unpublished data). The human-specific Yb lineage has members from only Yb8, Yb9 and the newly identified subfamilies. Each full-length human-specific Yb element was aligned against the reference genome using BLAST (Altschul *et al.*, 1990) with the e-value set to 10^{-5}. Based on the BLAST results, any insertions that match more than one genomic region with equal matching quality were omitted from further analysis as the source copy of these insertions could not be determined. The remaining sequences were divided into clusters based on their similarity with one another. The evolutionary relation between members of each cluster was obtained by constructing a phylogenetic tree using the neighbour joining method rooted with the Yb8 consensus sequence, and some cases

were supplemented with network analysis using the median joining method (Bandelt *et al.*, 1999).

2.2.2 Identification of new Alu Yb subfamilies

Position information for all Alu Yb8 and Yb9 elements from the latest major version of the human genome assembly GRCh37 were retrieved from RepeatMasker (Smit *et al.*, 2010) and the sequence for each insertion was retrieved from the reference genome. The poly-A segments from both the 3' end and the middle were removed manually. The pairwise alignment for all Yb9 sequences was visualized in MEGA5 (Tamura *et al.*, 2011). A signatory sequence was constructed encompassing each of the signature insertions at the 201st position and the mutation at the 259th position. The sequences were conserved across all Alu Yb insertions except for the mutation/insertion base. These sequences were aligned against the reference genome using BLAST with an e-value of 10^{-5}. The resulting matches were filtered using an in-house Perl script to retain only the sequences that have the signature mutation/insertion. To identify additional insertions of the new subfamilies that are absent in the reference genome, genome sequencing and alignment data from the 1000 Genome Project were downloaded to our local server. New insertions for Alu Yb8 and Yb9 in the six high coverage genome datasets from phase 1 of the 1000 Genome Project were identified in a separate study (Luo *et al.*, 2011); the read cluster for each predicted novel insertion contains all reads from the inserted region. From the mobile element insertion list generated from the pilot phase 1 data of the 1000 Genome Project (Stewart *et al.*, 2011), we collected 304 Alu Yb8 and Yb9 insertions that are absent in the reference genome but were detected in one or more of the test genomes for which a complete insertion sequence could be constructed. A custom BLAST

database was created to contain all these new insertion sequences, and the signature sequences were aligned against this custom database using the abovementioned criteria.

2.2.3 Validation of Yb11 insertions outside the reference genome

The insertion of T after the 200th nucleotide in Yb11 can be the result of a sequencing error since the preceding base is also a T. To eliminate the possibility of erroneous results, all reads sequenced by Sanger's method were downloaded from the NCBI trace database to our local server. The Yb11 signatory sequence was aligned against these reads to identify the reads that contain Yb11. A total of 130 reads were found to contain the Yb11-specific T insertion. The Phred quality score of the site of the T insertion in each read was analysed using a custom Perl script. Three out of fifteen loci could be confirmed using these trace data. Of the remaining twelve Yb11 insertions that are outside the reference genome sequence, primers could be designed for six Alu insertions. Five insertions could be amplified by PCR in DNA samples NA19239 and NA19240 from the Coriell Cell Repositories (http://ccr.coriell.org) and an in-house mixed DNA, all of which received approval from the Brock University Research Ethic Board. The amplified products were sequenced using the Sanger method at The Centre for Applied Genomics. The sequencing primers include locus-specific flanking primers and two Alu-internal primers designed from the 5' and 3' ends of the Yb11 consensus sequence, which are TGGCTCACGCCTGTAATC and GACGGAGTCTCGCTCTGTC, respectively. The internal primers help with difficulties in sequencing through the poly-A regions within Alu sequences. The sequences were aligned using clustalW to analyse the Yb11-specific site. All new Alu insertion sequences not covered by dbRIP were processed for deposition into dbRIP (http://dbrip.brocku.ca) under the study ID 2013-02.

2.2.4 Analyses of the Yb8a1, Yb10 and Yb11 insertion polymorphisms and evolution relations

To assess the level of polymorphism among the insertions of the three new subfamilies, the start and end position of each insertion was compared with structural variation (Mills *et al.*, 2011) and mobile element insertion (Stewart *et al.*, 2011) data from the 1000 Genome Project and with entries from dbRIP (Wang *et al.*, 2006b). The phylogenetic tree for all full-length Alu Yb9, Yb8a1, Yb10 and Yb11 insertions along with the putative source Yb8 copies obtained from previously mentioned clusters was constructed using the neighbour joining method (Saitou & Nei, 1987). All alignments and phylogenetic trees were visualized using the MEGA software (Tamura *et al.*, 2011). The evolutionary distance and sequence divergence within and between subfamilies were calculated using the maximum composite likelihood model (Tamura *et al.*, 2004) involving 181 full-length Yb9, 65 Yb8a1, 8 Yb10 and 15 Yb11 nucleotide sequences without poly-A sequences at the 3' end and in the middle.

2.3 Results and Discussion

2.3.1 Evolution of recent AluYb elements

Of all Yb copies found in the human genome, 80% (2,545 of 3,179) are identified as human-specific (hsYb), that is, they became integrated into the human genome after the human–chimp divergence, and they only include members of the Yb8 and Yb9 subfamilies (Tang et al., manuscript in preparation). In this study, we included all full-length hsYb elements in an attempt to assess their evolutionary pattern and backtrack

their putative source genes. All such hsYb elements were aligned against all Yb7, 8 and 9 sequences in the reference genome to group similar sequences into clusters. For each cluster, a phylogenetic tree was constructed with an outgroup subfamily consensus sequence as its root to assess the evolutionary relation among clusters and members of each cluster. The phylogenetic topology for each cluster can provide information on the potential parent copy for other members in the cluster. In an analysis involving only hsYb8 elements and their best matches, one particular cluster consists of 714 Yb8 elements. The phylogenetic tree involving all of these elements indicates that one copy of Yb8 (at hg19/chr10:10493416-10493732) seemed to have generated multiple active Yb8 copies that further retro-transposed to produce eventually 713 copies or 54% of all 1,322 hsYb8 elements studied (Figure 2.1). This master Yb8 element was most likely the major driver of the Yb8 expansion after the human–chimp divergence. Eight other Yb8 elements were detected that generated at least ten copies of offspring Yb8 elements. These Yb8 elements with lower activity level comply with the 'stealth driver' model of Alu evolution, which states that the stealth drivers do not generate as many copies of Alu as the master gene does, but rather function primarily to maintain the genomic retrotransposition capacity over a period of time (Han *et al.*, 2005).

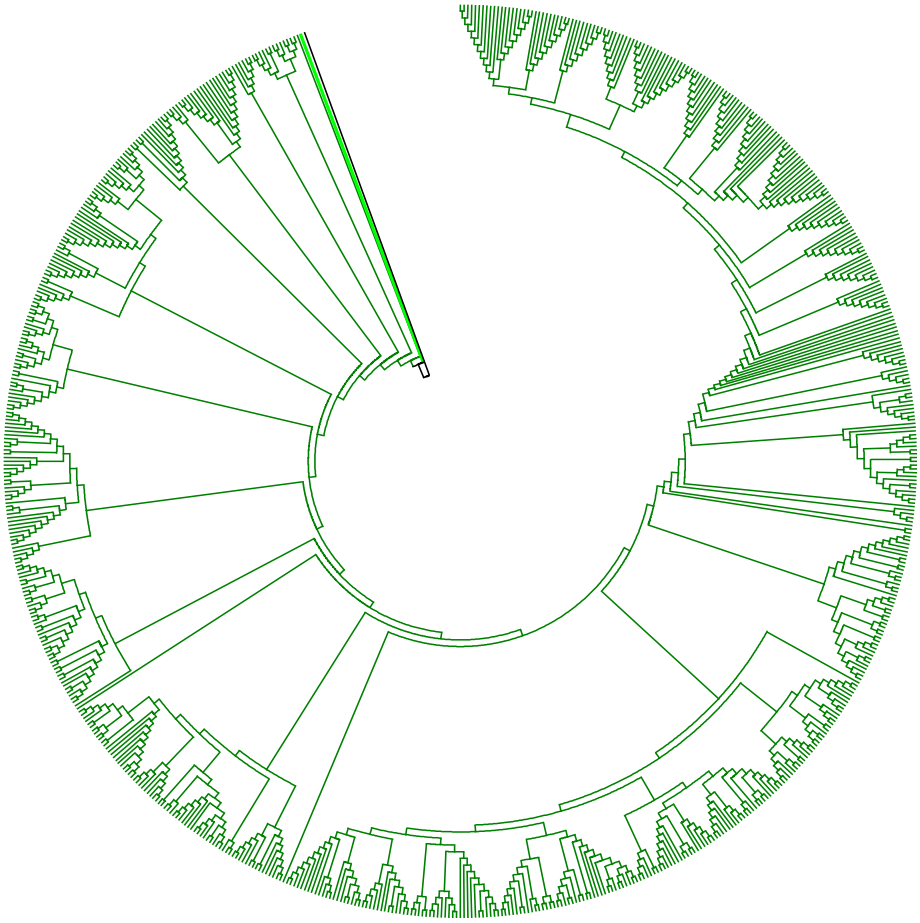


Figure 2.1 Cladogram with 714 hsYb8 elements constructed by the neighbour joining method. The element marked with a bold line (at hg19/chr10:10493416-10493732) is likely to be the source copy of all others in the tree. The tree was rooted using the Yb8 consensus, which is indicated by the black line.

A similar approach was taken to backtrack the evolutionary pathway of hsYb9 elements, involving identification and clustering of best-matched sequences from the whole genome. While almost all of the Yb9 elements tested aligned best with one another, 16 elements aligned best with 16 different Yb8 elements. When a phylogenetic tree was constructed with all hsYb9 elements and these 16 Yb8 elements, one particular Yb8 element at chr14:101990881-101991202 was found to be the source of all the hsYb9

elements, having generated multiple active Yb9 elements that subsequently generated 131 additional full-length hsYb9 copies (Figure 2.2). Along the evolutionary path of hsYb9, shown in Figure 2.2, some clusters have Yb8 elements, which may have resulted from either reverse mutation to produce Yb8 elements, or gene conversion or misannotation of Yb9 copies as Yb8 (Roy *et al.*, 2000).

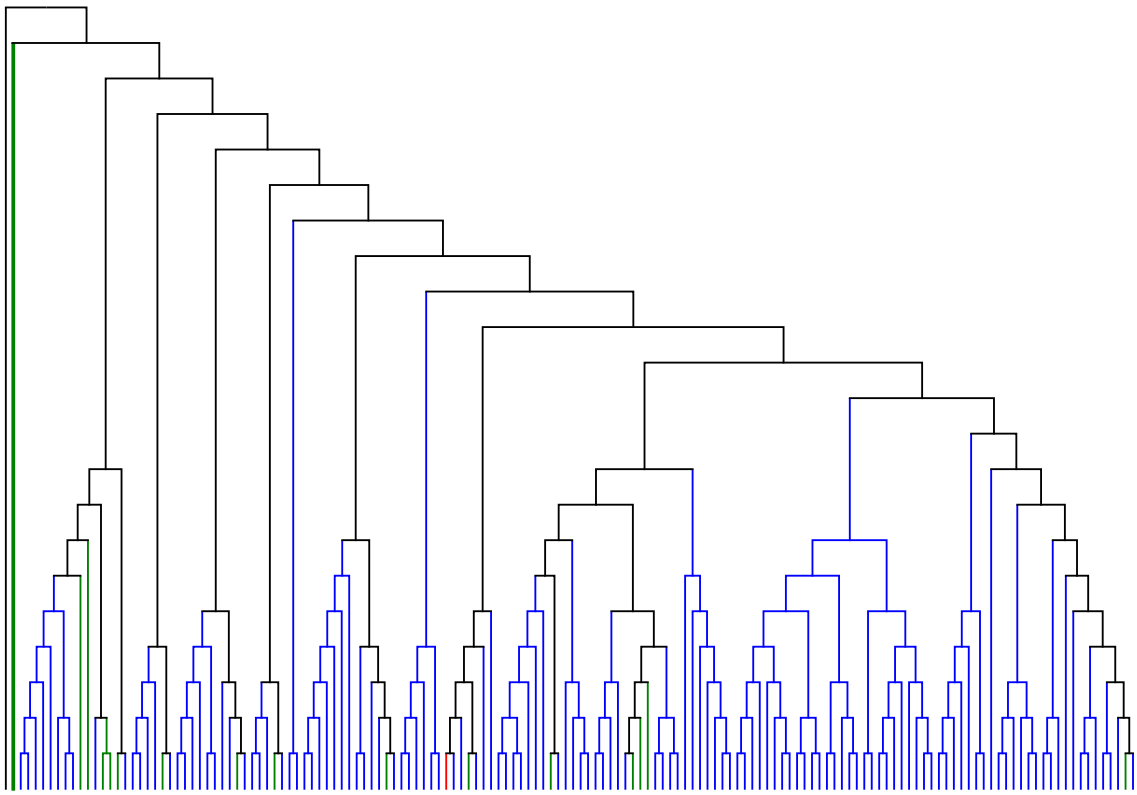


Figure 2.2 Cladogram with 131 hsYb9 and 16 Yb8 elements constructed by the neighbour joining method. Alu Yb9 and Yb8 elements are shown in blue and green, respectively. There is one Yb8a1 element in the cluster that matches best with one of the Yb9 elements, shown in red. The Yb8 copy shown in bold green is likely to be the source of all Yb9 copies in the cladogram. The Yb8 consensus (root) is shown in black.

2.3.2 Identification of novel Alu Yb subclasses

Different subfamilies of the Yb lineage are characterized by specific mutations, and the subfamilies are defined according to the number of mutation sites with respect to the Alu Y consensus sequence (Batzer *et al.*, 1996). Identification of new subfamilies is basically the identification of a set of Alu elements that share a particular mutation at a specific site that has not been previously reported. Using a computational approach, we performed a genome-wide analysis of Alu elements that are currently annotated as Yb8 and Yb9, the two most recent subfamilies of the Yb lineage known to date, to investigate whether any specific mutation beyond the Yb8 and Yb9 signature mutations is shared by more than one element. To do so, a set of full-length members of the Alu Yb8 and Yb9 subfamilies were retrieved from the latest assembly of the human reference genome sequence GRCh37, and multiple sequence alignment was performed after the poly-A segments were removed. Upon careful examination of the alignment data, two specific mutations were observed in multiple Yb9 and Yb8 elements at the 201st (insertion of T) and 259th (G → A) positions, respectively. We also observed that Alu sequences with the single base insertion after the 200th position always carry the mutation at the 259th position and the Yb9 diagnostic mutation at the 174th position, but not all sequences with a mutation at the 259th position contain the other two mutations. This is only possible if the sequences with the $259^{G\rightarrow A}$ mutation originated from the Yb8 subfamily as the first event and then a subset of these sequences accumulated the Yb9-diagnostic $174^{C\rightarrow G}$ mutation, or vice versa, giving rise to another new subfamily, which subsequently accumulated the 200^{+T} insertion to generate yet another subclass of Yb elements. Following the standard nomenclature of Alus (Batzer *et al.*, 1996), we named the sequences with the $259^{G\rightarrow A}$ mutation Alu Yb8a1, the sequences with the $259^{G\rightarrow A}$ and

$174^{C\rightarrow G}$ mutations Alu Yb10, and the sequences with the $259^{G\rightarrow A}$ and $174^{C\rightarrow G}$ mutations and the 200^{+T} insertion Alu Yb11 (Figure 2.3). When a Yb8a1 signatory sequence of 30 bases was constructed and aligned against the human reference genome, 99 Yb10 copies were identified, among which 75 copies did not have the $174^{C\rightarrow G}$ mutation (Yb8a1), 8 had the $174^{C\rightarrow G}$ mutation (Yb10), and 16 copies had both the $174^{C\rightarrow G}$ mutation and the 200^{+T} insertion (Yb11). A 24-nucleotide-long signatory sequence was also constructed for Yb11, and when this sequence was aligned against the reference genome, 16 matches were detected, all of which overlap with the results from the Yb10 signatory sequence-whole genome alignment, which provides evidence for the accuracy of the method. In the end, we were able to detect 75 Yb8a1, 8 Yb10 and 16 Yb11 insertions in the reference genome (Appendix I: Table 1).

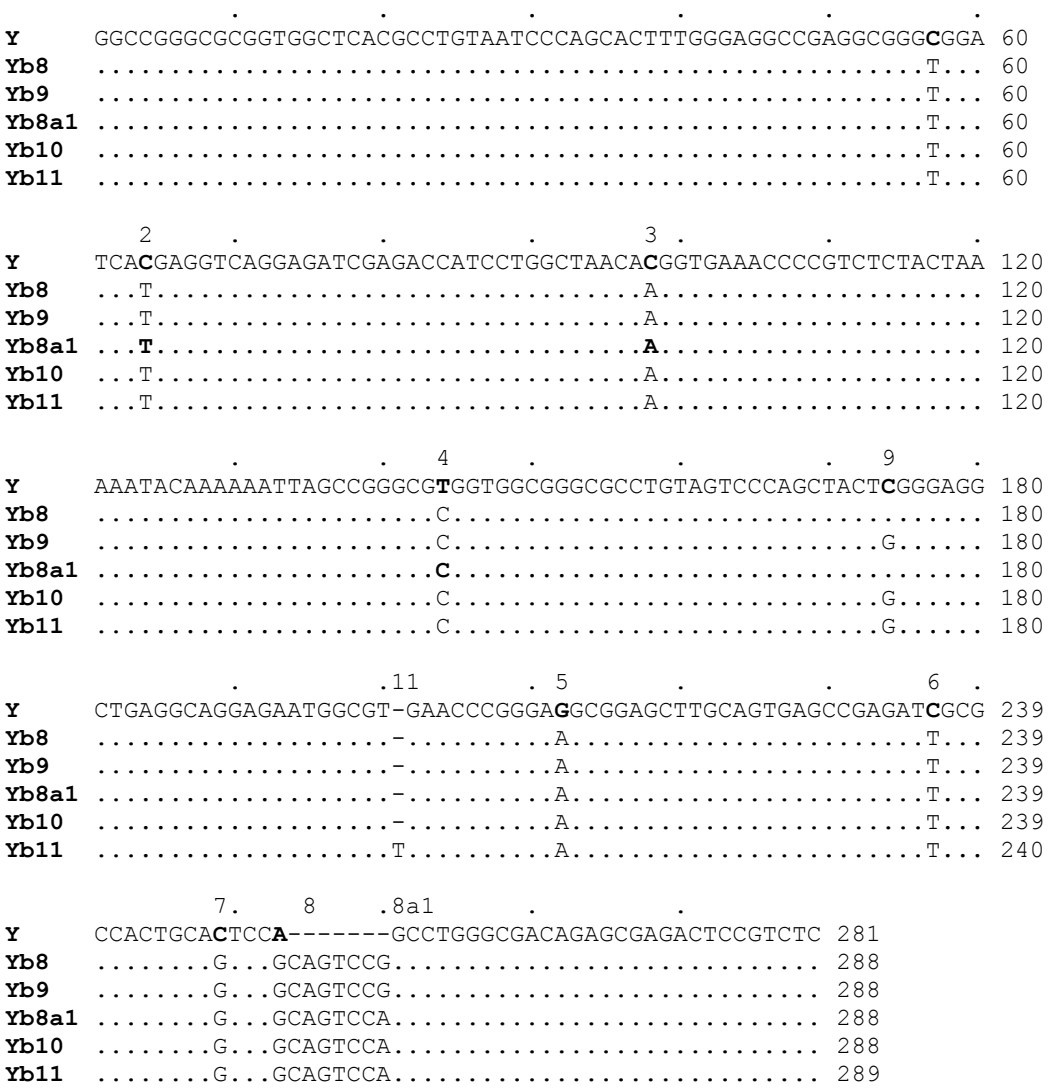


Figure 2.3 Consensus sequences of Alu Y, Yb8, Yb9, Yb8a1, Yb10 and Yb11. The signatory mutations are numbered in chronological order as they were identified using Alu Y as the baseline.

Besides the reference genome, we also analysed 1000 Genome Project (1KGP) data and sequencing trace data from HuRef (Levy *et al.*, 2007), to identify insertions of the newly identified subfamily members that are absent in the reference genome. We collected all of the Yb8 and Yb9 insertions that are absent from the reference genome but

present in one or more individual genome sequences in the 1KGP data, for which sufficient insertion sequences could be constructed. Signature sequences for Yb8a1, Yb10 and Yb11 were then aligned against these sequences and the HuRef sequencing, resulting in the detection of an additional 6 Yb8a1, 3 Yb10 and 15 Yb11 insertions outside the reference genome. The insertion of T in the Yb11 elements outside the reference genome was confirmed by PCR amplification and sequencing for five of these 15 loci and by manually checking the sequencing data from the National Center for Biotechnology Information (NCBI) trace database for three of them (Figure 2.4; Figure 2.5; Appendix I: Table 2). Therefore, we were able to identify a total of 81 Yb8a1, 11 Yb10 and 31 Yb11 insertions, and we can expect that more of these will be identified after processing more personal genomes.

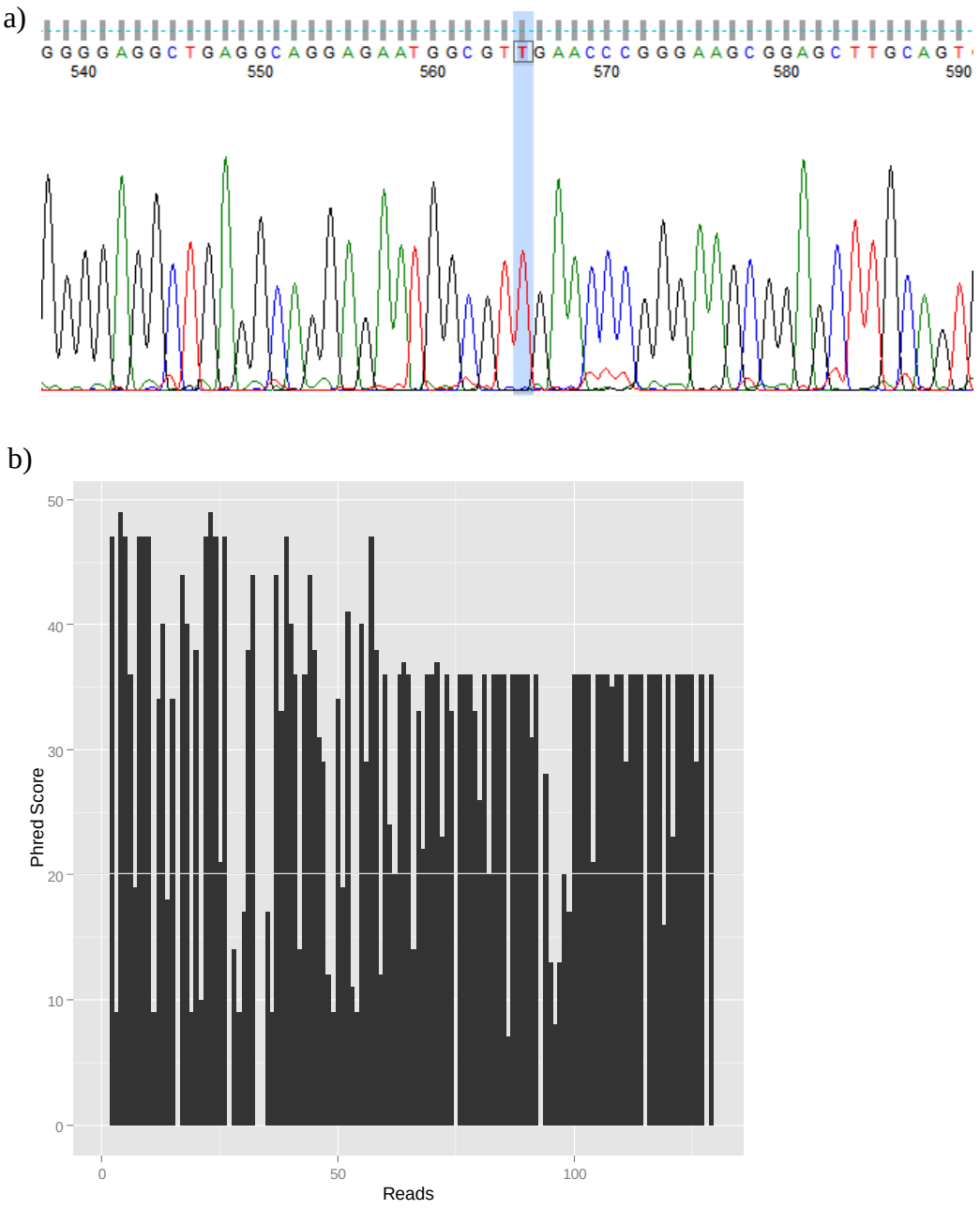


Figure 2.4 Identification and quality analysis of Yb11-specific insertion of T in human genome fragments sequenced by Sanger's method obtained from NCBI database. a) Chromatograph of a sequence read output from Sanger's method (TI: 1747216562). The Yb11-specific insertion of T is highlighted. The top bars above the nucleotide labels represent the Phred quality scores for individual bases. b) The Phred quality score of Yb11-specific insertion site in all reads that have the Yb11 sequence. Each bar represents the site of T-insertion in each individual sequence read. A Phred score of 10 denotes 90% base call accuracy, 20 denotes 99% accuracy and 50 denotes 99.999% accuracy. A Phred score of 0 indicates that the base could not be identified due to poor sequencing quality.

```
3528_predicted      AATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGG 60
108507              AATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGG 60
3528                AATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGG 60
128385              AATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGAGGYTGAGGCAGG 60
56065               AATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGG 60
55925               AATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGG 60
                    ************* ************************************ *********

3528_predicted      AGAATGGCGTTGAACCCGGGAGGCGGAGCTTGCA 94
108507              AGAATGGCGTTGAACCCGGGAAGCGGAGCTTGCA 94
3528                AGAATGGCGTTGAACCCGGGAAGCGGAGCTTGCA 94
128385              AGAATGGCGTTGAACCCGGGAAGCGGAGCTTGCA 94
56065               AGAATGGCGTTGAACCCGGGAAGCGGAGCTWGCA 94
55925               AGAATGGCGTTGAACCCGGGAAGCGGAGCTTGCA 94
                    ********************* ******** ***
```

Figure 2.5 Alignment of the partial sequences of amplified Yb11 loci that are absent in the reference genome but present in one or more other individual genome sequences. Sequencing was done using Sanger's method and all bases used in this alignment were called with Phred quality score of above 20. The Yb11-specific insertion of T is highlighted. The amplified loci are with IDs of 55925, 3528, 128385, 56065 and 108507. "3528_predicted" is the predicted sequence obtained from the 1000 Genome Project data for the locus ID P1_MEI_3528&P2_MEI_466.

2.3.3 Age estimation

Mutation densities were calculated for each subfamily to estimate the approximate age of the new subfamilies. Only full-length or near full-length Alu elements in the reference genome were considered (65 Yb8a1 out of 75, 8 Yb10, and 15 Yb11 out of 16) and the poly-A regions in the middle and at the end were removed. For the 65 elements from the Yb8a1 subfamily, the non-CpG mutation density was 0.29% (43 out of 14,625 total non-CpG bases). Using a neutral rate of evolution of 0.15% per million years for primate intervening DNA sequences (Miyamoto *et al.*, 1987) along with the non-CpG mutation density, the average age of the Yb8a1 subfamily was estimated to be 1.93 million years old. For the 8 Yb10 elements, 5 non-CpG mutations were detected out of a total of 1,904 non-CpG nucleotides constituting only 0.26% of them, indicating an estimated age of 1.73 million years for Yb10. For the Yb11 subfamily, 15 elements were

analysed with a total of 3,720 non-CpG nucleotides; only 4 of these had mutated, yielding a neutral mutation density of 0.107% and an estimated age of 0.71 million years. To assess how recent these subfamilies are in relation to the already known Yb subfamilies, the age of Yb9 was also estimated. A total of 166 non-CpG mutations were identified from 254 Alu Yb9 family members containing 51,562 non-CpG nucleotides; 73 members were not included in the calculations due to a 5' truncation or a large deletion inside the Yb9 element. Using the same neutral rate of evolution and the non-CpG mutation density of 0.32% (166/51,562), the average age of the Yb9 subfamily members was estimated to be 2.15 million years. The age of the Yb9 subfamily estimated in this study is much older than that estimated initially by Roy-Engel *et al*. (Roy-Engel *et al.*, 2001), mainly because the total number of Yb9 elements in their study was much smaller than in this study. However, our estimation of the age of Yb9 is very close to that identified in a similar study, which estimated the age of Yb9 as 2.32 million years (Carter *et al.*, 2004). The estimated age for Yb8a1 indicates that this subfamily originated almost at the same time as Yb9, which is evidence that Yb8a1 originated from Yb8. The Yb10 subfamily, which evolved 1.73 million years ago, should be mostly fixed across all human populations, while the Yb11 subfamily, at only 0.71 million years old, is most likely to be highly polymorphic among human populations because it is the youngest. The level of polymorphism for these newly identified subfamilies with respect to their ages are examined further in the following section.

2.3.4 Level of polymorphism

The Alu Y family is evolutionarily the 'youngest' Alu family and the Yb lineage was found to be one of the largest and most active lineages of all young Alu elements (Carter *et al.*, 2004; Jurka, 1993; Wang *et al.*, 2006a). Out of the 2,433 full-length Yb elements found in the human genome, 499 were found to be polymorphic for their presence or absence between individuals and/or populations, and a further 304 Yb copies were identified in individual genome sequences that are not present in the reference genome (Jurka *et al.*, 2005; Stewart *et al.*, 2011). Since the majority of Yb elements became inserted into the human genome 3 to 4 million years ago, we suspect that the very recently evolved subfamilies contribute most to the polymorphism due to the Yb lineage since the divergence of the various human populations from their common ancestor occurred only 100,000 years ago (Carter *et al.*, 2004). We assessed the level of polymorphism for all identified Yb8a1, Yb10 and Yb11 insertions by surveying Alu insertions and deletions in personal genomics data. We compared the insertions that are present in the reference genome with the structural variation data from the 1000 Genome Project (Mills *et al.*, 2011). Of these, 13 out of 16 (approximately 81%) Yb11 elements and 2 out of 8 (25%) Yb10 were found to be dimorphic, while 22 out of 75 (approximately 29%) Yb8a1 present in the reference genome are polymorphic. We then compared these polymorphic insertions with dbRIP to identify how many of them have previously been reported as polymorphic and found that 7 and 2 polymorphic Yb8a1 and Yb11 elements, respectively, overlap with dbRIP data (Wang *et al.*, 2006b). Combining insertions both inside and outside the reference genome, a total of 28 out of 31 (approximately 90%) Yb11 and 5 out of 11 (approximately 45%) Yb10 were found to be polymorphic, while only 28 out of 81 (approximately 34%) of Yb8a1 insertions were

identified as polymorphic. The difference in the level of polymorphism is inversely related to the age of the lineage, that is, the higher the polymorphism level among individuals and/or populations, the more evolutionarily recent the lineage. The difference in the fraction of polymorphic members among the three novel subfamilies confirms that Yb11 has evolved more recently than Yb10 and Yb8a1. The relative newness of the Yb11 lineage is further substantiated when we looked at the sequence divergence within the members of each subfamily (Table 2.1). The mean evolutionary divergence between each pair of sequences in the Yb8a1, Yb9, Yb10 and Yb11 subfamilies was estimated to be 0.016, 0.026, 0.015 and 0.006, respectively. The divergence value is directly related to the age of the population, that is, the older the set of sequences, the more evolutionarily divergent the sequences are. The mean divergence values provide another line of data suggesting that Yb8a1, Yb10 and Yb11 evolved chronologically during the evolution of humans.

Table 2.1 Estimates of evolutionary divergence between and within full-length *Alu* Yb9, Yb10 and Yb11 elements.

	Alu Yb8a1	**Alu Yb9**	**Alu Yb10**	**Alu Yb11**
Alu Yb8a1	0.016[a]			
Alu Yb9	0.026[b]	0.026		
Alu Yb10	0.019	0.022	0.015	
Alu Yb11	0.015	0.017	0.011	0.006

[a]The average of base substitutions per site of all pairwise comparisons within the group.
[b]The average of base substitutions per site of all pairwise comparisons among the members of the two groups compared.

We also examined the distribution of all polymorphic members of Yb8a1, Yb10 and Yb11 in Yoruban, European, Chinese and Japanese populations using the data from the 1000 Genome Project (Stewart *et al.*, 2011). It was observed that 50%, 64% and 59% of

polymorphic elements are present in the Yoruban population for the Yb8a1, Yb10 and Yb11 subfamilies, respectively (Figure 2.6). These numbers are higher than the equivalent numbers for the other non-African populations examined. The highest number of polymorphic elements were expected to be present in the Yoruban population as this was the oldest population tested in this study (Stringer & Andrews, 1988). While the presence or absence of some of the polymorphic elements could not be ascertained for the Chinese and Japanese populations (they are flagged as 'unascertained'), the majority of the rest (approximately 66%) were present in one or both of the Asian populations. Among these, only one Yb8a1 insertion was found to be specific to the Chinese population and the rest are all shared by one or more other populations. In contrast, 15 Yb8a1, 5 Yb10 and 10 Yb11 insertions are specific to the Yoruban population, and 2, 3 and 4 of each of Yb8a1, Yb10 and Yb11 insertions are specific to the European population. This suggests that the number of population-specific insertions decreases with the age of the population. In other words, the older the population, the more time there has been for active young Alu elements to retrotranspose, creating a direct relation between the number of population-specific Alus and the age of population.

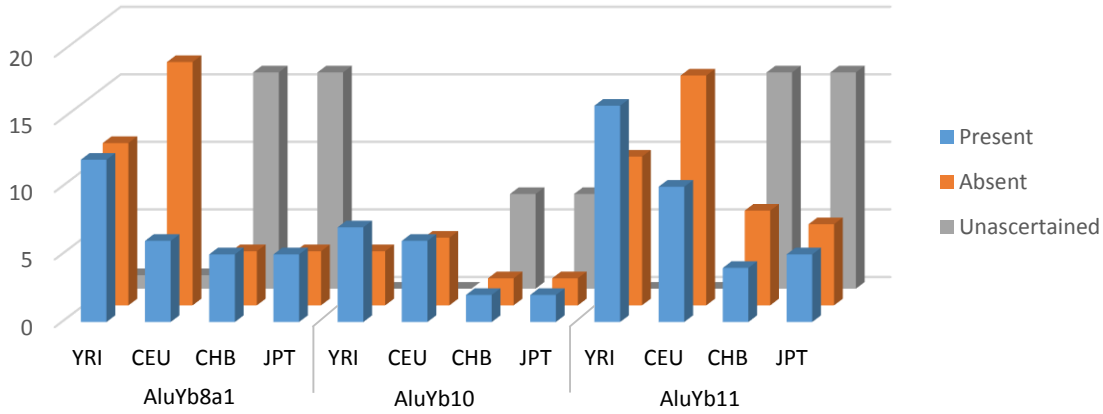


Figure 2.6 The level of polymorphism for the Yb8a1, Yb10 and Yb11 subfamilies. The blue columns at the front indicate the number of polymorphic insertions observed in the population and the orange columns in the middle represent the number of insertions observed in other populations but not in the population. The presence or absence of polymorphic insertions in Chinese and Japanese populations could not be determined and these are labeled as 'unascertained' and represented by grey bars. CEU, Residents of Utah with European ancestry; CHB, Chinese from Beijing; JPT, Japanese; YRI, Yoruban population.

2.3.5 Evolutionary pathways for the three new Alu Yb subfamilies

The master gene model is the most widely accepted model for the generation of new Alu subfamilies (Deininger *et al.*, 1992) even though there many doubts about the details of this model (Batzer *et al.*, 1995b; Cordaux *et al.*, 2004; Price *et al.*, 2004; Schmid, 1993). While this model only gives a hierarchical evolution for the different subfamilies, the specific evolutionary pathways for the generation of different Yb lineages have yet to be characterized. The evolution of Yb9, Yb8 and Yb7, the three most recent and abundant subfamilies of the Yb lineage, occurred sequentially (Roy-Engel *et al.*, 2001).

In our study, we predict that the evolution of Yb11 took a strict sequential linear pathway from Yb10 since it contains one more mutation than Yb10 diagnostic mutations, while the Yb10 subfamily evolved from either Yb8a1 or Yb9 following one or more pathways (Figure 2.7). A tree using the neighbour joining method was constructed among 25, 181, 65, 8 and 15 full-length Yb8, Yb9, Yb8a1, Yb10 and Yb11 elements, respectively, rooted with the Yb8 consensus sequence (Figure 2.8). The 25 Yb8 elements were included because these are the only Yb8 copies that one or more of Yb9, Yb8a1, Yb10 and Yb11 had the best similarity score with. It was observed from the topology that 77% of all Yb8a1 elements have evolved from one individual Yb8 copy, and 63 out of 65 Yb8a1 copies tested are evolutionarily closest to members of the Yb8 subfamily. This supports the hypothesis that Yb8a1 evolved from Yb8 as a separate lineage from Yb9. Among the 15 Yb11 copies included in the phylogenetic analysis, all of them have common nodes with copies from Yb10 elements, supporting their linear evolutionary pathway from the Yb10 subfamily.

The diagnostic mutations of the Yb10 subfamily are predicted to have evolved by following one of two pathways: (1) a Yb9 element obtained the Yb8a1-specific mutation and retrotransposed to generate the Yb10 subfamily or (2) a Yb8a1 element obtained the Yb9-specific mutation subsequently generating the Yb10 subfamily. The phylogenetic analysis on its own does seem to favour the latter option since the major branch leading to the Yb10/Yb11 lineage is closer to the Yb8a1 cluster. For additional evidence, an evolution network was constructed for all full-length members of the four subfamilies of interest using the median joining method (Bandelt *et al.*, 1999). The network shows that the majority of the Yb10 elements are linked closer to multiple Yb8a1 elements than to

Yb9 (Figure 2.9), further supporting the prediction that the evolution of Yb10 was from Yb8a1 by gaining the Yb9 mutation. The accumulation of the Yb9-specific mutation in the Yb8a1 copy parent to create the Yb10 subfamily may have occurred by gene conversion and requires further analysis for confirmation. A second line of evidence for the evolutionary pathway proposed here is provided by the linear pairwise evolutionary distances calculated for the Yb9, Yb8a1, Yb10 and Yb11 elements (Table 2.1). The mean evolutionary distance for all sequences between Yb10 and Yb11 was calculated as 0.011, which is lower than the distance between Yb9 and Yb11 (0.017) or Yb8a1 and Yb11 (0.015) indicating the sequential evolution of Yb11 from Yb10 and with Yb8a1 being closer than Yb9 to Yb11.

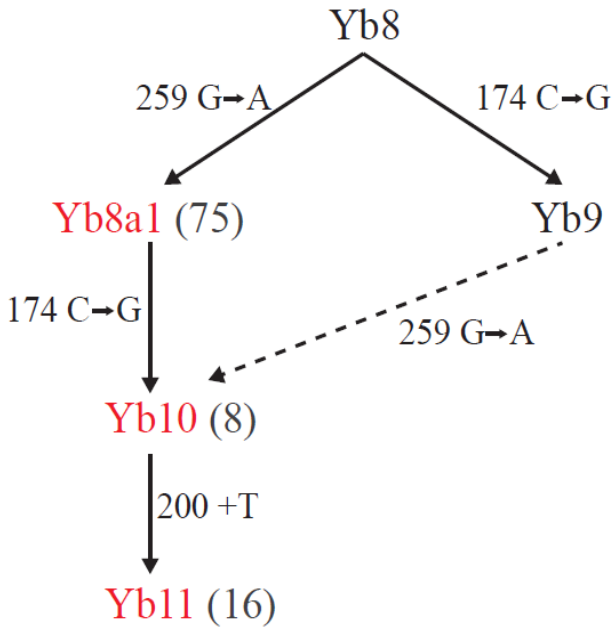


Figure 2.7 Evolution of the recent AluYb lineage. The subfamilies in black are the current known subfamilies and the subfamilies in red are novel and proposed in this study. The numbers accompanying each subfamily are the total number of copies found in the human reference genome. The dotted line is the less convincing alternative pathway for the evolution of the Yb10 subfamily.

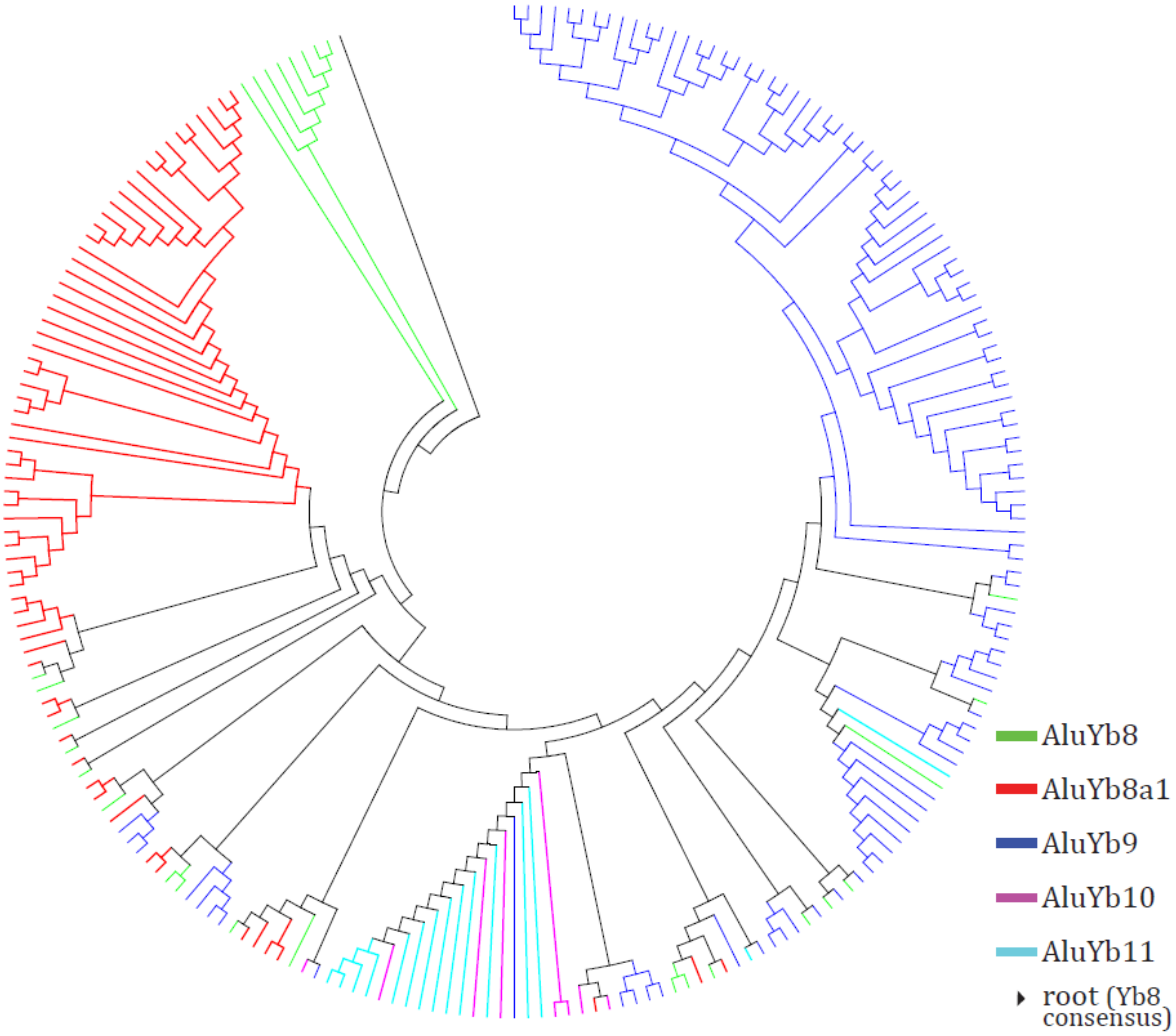


Figure 2.8 Cladogram of all full-length Yb9, Yb8a1, Yb10, and Yb11 elements using the neighbour joining method. The tree is rooted with the Alu Yb8 consensus sequence, which is shown in black at the top left.

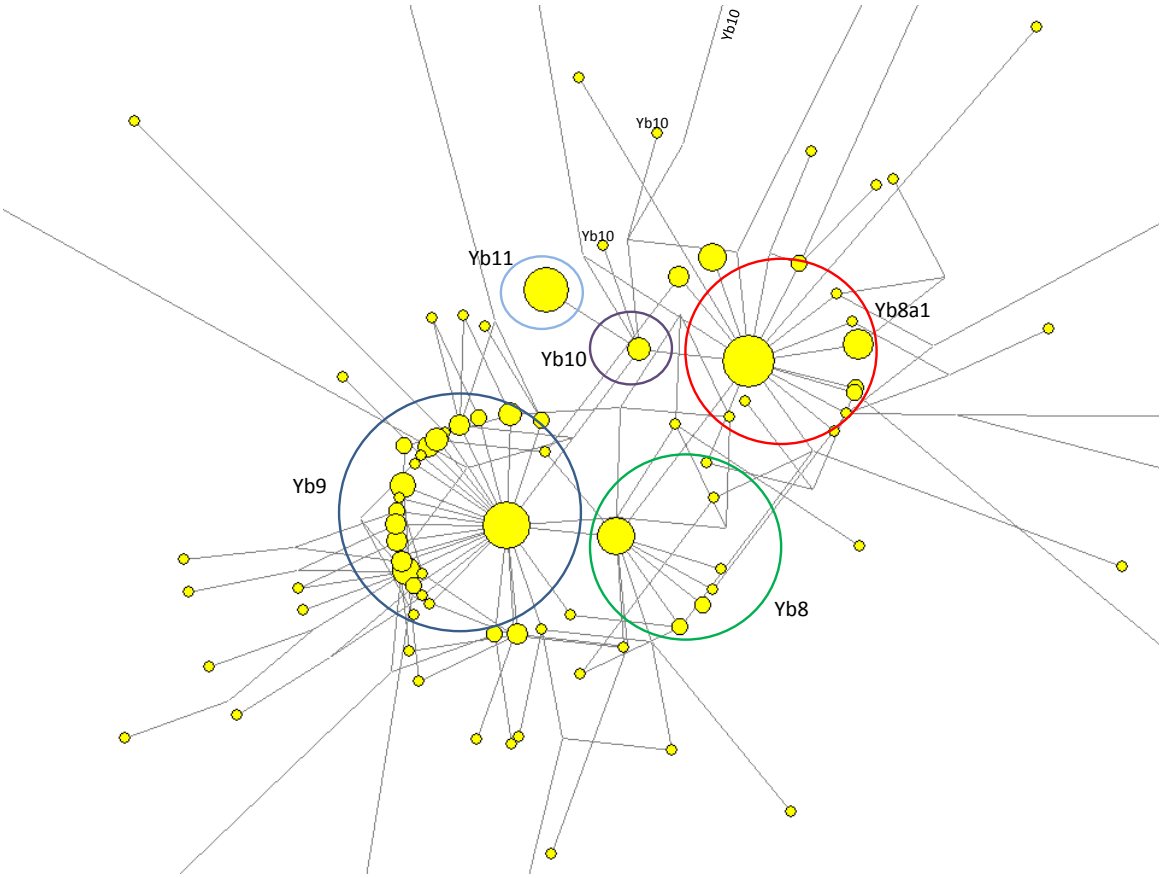


Figure 2.9 Network between full length Alu Yb8, Yb9, Yb8a1, Yb10 and Yb11 elements using Median Joining method. Each subfamily forms a cluster based on sequences that is annotated by circles. The length of each connecting line is relative to the number of mutations. The novel Alu subfamily Yb8a1 is closer to Yb8 cluster, Yb10 is closer to Yb8a1 and Yb11 has connection only with Yb10 members.

Each of the Yb8a1, Yb10 and Yb11 subfamilies was also tested using the molecular clock (ML) to assess if all full-length members in each subfamily evolved at a homogeneous rate. A maximum likelihood test of the ML hypothesis was performed separately for each of the Yb8a1, Yb10 and Yb11 phylogenetic tree topologies and sequence alignments (Felsenstein, 1981). The ML hypothesis states that all tips of the tree should be equidistant from the root of the tree, or in other words the rate of evolution of all branches in the tree is uniform. The maximum likelihood, $-\ln L$, was calculated to

be 990.971 and 907.158 for with-clock and without-clock phylogeny, respectively, for Yb8a1, 466.906 and 455.855 for with-clock and without-clock phylogeny, respectively, for Yb10, and 481.574 and 474.459 for with-clock and without-clock phylogeny, respectively, for Yb11. The chi-square test based on the difference in the likelihood ratio between with-clock and without-clock phylogeny rejected the null hypothesis of uniform evolution for both Alu Yb8a1 and Yb10 insertions at a 5% significance level with $P < 0.0001$ and $P < 0.001$ for Yb8a1 and Yb10, respectively. However, we failed to reject the null hypothesis of an equal evolutionary rate for all Yb11 insertions at a 5% significance level ($P < 0.43$). This indicates that neither the Yb8a1 nor the Yb10 subfamily evolved at a uniform evolutionary rate, and that the evolution of the subfamily Yb11 has been uniform. This provides further evidence that the Yb8a1 and Yb10 subfamilies are older than the Yb11 subfamily since evolutionary uniformity is more likely in a recently evolved lineage. Furthermore, when the evolutionary relations for all full-length Yb8a1, Yb9, Yb10 and Yb11 elements were analyzed, more divergence among members of Yb8a1 and Yb9 was observed than among the members of Yb10 or Yb11 (Figure 2.10), another indication that the former subfamilies are older than the latter.

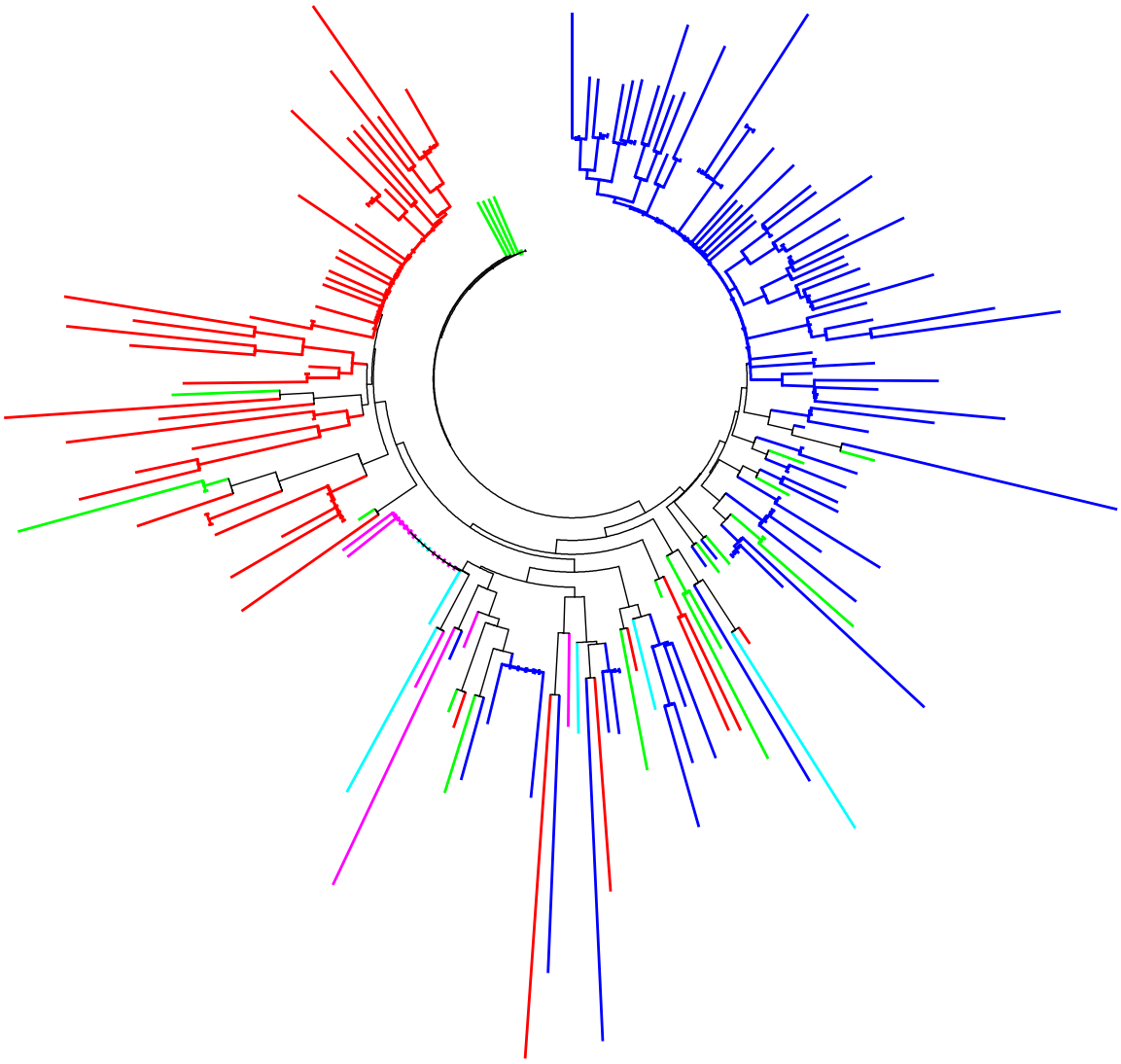


Figure 2.10 Evolutionary relationships of all full-length Yb9, Yb8a1, Yb10 and Yb11 elements. The green, blue, red, magenta and neon lines represent Yb8, Yb9, Yb8a1, Yb10 and Yb11 elements respectively. The tree is rooted with AluYb8 consensus sequence. The evolutionary history was inferred using the Neighbor-Joining method. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. All ambiguous positions were removed for each sequence pair. There were a total of 254 positions in the final dataset.

2.4 Conclusions

The Alu Yb lineage has an extended evolutionary history in the human genome. Even though the lineage evolved before the human–chimp divergence, most of the insertions occurred in the last 3 to 4 million years and some copies of this lineage still retain the ability to retrotranspose. One such active Yb8 copy has generated almost 60% of all human-specific Yb8 copies and several others have generated more than ten copies, indicating the presence of both a master copy and stealth drivers for this subset of Yb8 elements.

The tracking of the source copy in this study enabled us to identify the potential master gene of all Yb9 elements. The relatively higher activity of the Yb lineage than almost all other Alu lineages has generated several subfamilies that were previously undetected and which share a specific pattern of mutations. Three such novel subfamilies proposed in this study are Yb8a1, Yb10 and Yb11. Even though Yb8a1 and Yb10 are believed to have evolved within a short time of each other, only eight copies of Yb10 have been detected in the human reference genome compared to 75 copies of Yb8a1. Furthermore, Yb9 has been estimated to be only 0.22 million years older than Yb8a1, yet the number of Yb9 copies in the human genome is almost five times larger than the number of Yb8a1 copies. This indicates that not all of the Alu subfamilies grew at an equal rate and that some mutation patterns may accelerate the rate of transposition. This is further supported by the fact that the Yb11-specific insertional mutation in the Yb10 sequence has accelerated the rate of retrotransposition resulting in 16 copies of Yb11 since it first evolved 0.71 million years ago. The possibility that certain mutations

accelerate the rate of transposition and their mechanism should be the subject of further study.

Yb11 is the latest subfamily to have evolved in this lineage and it is highly polymorphic among different individuals and/or populations. The generation of these young subfamilies indicates that Alus are still evolving, and this provides some clues regarding the future trend of Alu activity in the human genome. The homoplasy-free nature of Alu insertions makes these very recent genetic variants a valuable resource in forensics and for studying modern human population genetics and migration patterns.

Chapter 3 : Transposable elements are a significant contributor to tandem repeats in the human genome

(The content of this chapter was mostly copied from the published article: "Ahmed M, Liang P: Transposable elements are a significant contributor to tandem repeats in the human genome. 2012. *Comp Funct Genomics* 2012: 947089" with some minor text edits.

The candidate is the main author of this article and was responsible for generating all of the data included in the article. The manuscript was drafted by the candidate and edited by the corresponding author, Dr. Liang, to its final form.)

3.1 Background

Over half of the human genome consists of repeat elements. The two types of repeat elements that are prevalent in the human genome are tandem repeats (TRs) of sequences ranging from a single base to mega bases and interspersed repeats that mainly include transposable elements (TEs). The tandem repeats are classified in three major classes based on the size of the repeated sequence - microsatellites for short repeat units (usually <10 bp), minisatellites for head-to-tail tandem repeat of longer units (>10 and <100 bp) and satellites for even larger units (>100bp). Among all types of tandem repeats, minisatellites and microsatellites have gained increasing attention over the past decade due to their contribution to intra-species genetic diversity and uses as genetic markers in population genetic studies. These repeat sequences are widespread in all eukaryotic genomes (Charlesworth, 1994) from yeast to mammals and often are highly polymorphic in populations of the same species. Consequently they are often used as a marker in numerous genotypic tests, e.g., in forensic fingerprinting (Jeffreys & Pena, 1993; Jeffreys *et al.*, 1985; Spurr *et al.*, 1994; Tamaki *et al.*, 1995), in population genetics (Armour *et al.*, 1996), and in monitoring of DNA damage induced by ionizing radiation (Bois & Jeffreys, 1999). Minisatellites lately have been of particular interest because their expansion has been implicated in alteration of gene expression often leading to diseases (Sutherland *et al.*, 1998). Origin and expansion of microsatellites have been well studied and the most widely accepted mechanism underlying microsatellites states that the initiation takes place by chance, and then they are expanded by slipped-strand mispairing (Levinson & Gutman, 1987). On the other hand, origin of minisatellites and satellites is very difficult to study, and even though a significant progress has been made in

understanding the expansion and contraction of such repeats, a number of major aspects are still unresolved (Bois, 2003). For expansion and contraction of longer repeats, several lines of evidence suggest gene conversion during meiosis as the major mutational force rather than replication slippage (Murray *et al.*, 1999; Richard & Paques, 2000). As for the direction of expansion, it has been found to be usually polar, i.e., addition of new repeat unit occurs only at one end (Jeffreys *et al.*, 1994).

While the expansion of longer sequences is well studied, the origin or initiation of such repeats is difficult to understand because it is very unlikely for duplication of such long repeats to initiate by chance. There are two models that attempt to explain the initiation of minisattelites/satellites. One model postulates slipped-strand mispairing at non-contiguous repeats when there is a pause during replication (Taylor & Breden, 2000). A key feature of this model is that expanded TR's terminal repeat unit should be "incomplete", i.e., shorter than other repeat units by a number of nucleotides. The second model postulates that when a long sequence is flanked by direct repeats of 5-10 bp, it can be duplicated by replication slippage or unequal crossing-over (Haber & Louis, 1998).

The other major class of repeats in the genome, transposable elements, are ubiquitous in both prokaryotes and eukaryotes. TEs can mutate genomes by transposing to new locations or by facilitating homology-based recombination due to their abundance in the genome. At least 44% of the entire human genome is composed of TEs that belong to at least 848 families or subfamilies (Mills *et al.*, 2007). Majority of the TEs in humans is contributed by two classes, L1 and Alu. When human genome was compared with chimpanzee genome, more than 10,000 species-specific insertions were identified, over 95% of which are contributed by L1, Alu or SVA (Hedges *et al.*, 2004; Mills *et al.*, 2006;

Wang *et al.*, 2006a; Watanabe *et al.*, 2004). SVA is a composite element that is derived from three other repeat elements: SINE-R, VNTR and Alu. A small number of human-specific TE insertions are also contributed by Human Endogenous Retrovirus-K (HERV-K) (Mills *et al.*, 2006). These human-specific TE insertions indicate that these TE families are/were active after the divergence of humans from chimps ~6 million years ago. Alu family has three large sub-families, AluJ, AluS and AluY, with their ages being considered very old, old, and young, respectively.

Even though the effects of TRs and TEs are well-studied and understood individually, there have not been many studies that investigated the relationship between these two classes of repeat sequences. To our knowledge, the first study linking tandem repeats and transposable elements was reported by Jurka and Gentles (Jurka & Gentles, 2006) in an attempt to identify the origin and diversification of minisatellites derived from Alu sequences. Their work demonstrates how Alu sequences can be tandemly repeated because of short direct repeats flanking the repeat arrays. Later Ames et al. (Ames *et al.*, 2008) also reported 111,847 TRs overlapping with interspersed repeat sequences in an attempt to compare between single-locus TRs and multi-locus TRs. They included microsatellites and all types of interspersed repeats but did not analyze the relationship between TRs and TEs any further. In the current study, we for the first time assessed the genome-wide contribution of TEs to the generation of minisatellites/satellites TRs, revealing that at least 7,276 TRs or 23% of all minisatellites/satellites were derived from TEs. We compared and identified the classes of TEs that are more prone for generating TRs, and we also examined the mechanisms for initiation and expansion of the tandem repetition of the TEs.

3.2 Materials and methods

3.2.1 Collection of TR and TE data in the human genome

The Tandem Repeat data were downloaded to our local server from the Tandem Repeat Database (TRDB) (http://tandem.bu.edu/cgi-bin/trdb/trdb.exe) that documents the genomic positions of each repeat, consensus repeat sequence and number of repeats among an array of useful information (Gelfand *et al.*, 2007). The consensus sequences of all families and subfamilies of TEs were downloaded from RepBase (http://www.repbase.org) (Jurka *et al.*, 2005). The positions of all individual TEs in the human genome were downloaded from UCSC Genome Annotation Database for genome version hg19 (http://hgdownload.cse.ucsc.edu). The UCSC hg19 (NCBI Build 37) version of human genome sequence was downloaded from UCSC website and was compiled to create a database for BLAST. Algorithms to perform all analytic tasks were developed in-house using the programming language Perl on Unix platform.

3.2.2 Identification of TE-derived TRs

Output from TRDB for all TRs in the human genome was filtered using an in-house Perl script such that they meet the following criteria: repeat unit length ≥ 20bp, GC content ≥ 40%, repeat number ≥ 2 and sequence similarity among the repeat units in an array ≥95%. Many satellites are parts of a larger satellites which causes redundancy in the final set; to avoid this, overlapping TR arrays are separated and the TR with smallest period from each set of overlapping arrays were used for the subsequent analyses. A TR is considered to be derived from a TE if it meets one of the following two criteria: 1) the TR repeat unit sequences have a minimum of 70% similarity with the consensus sequence of a human TE; 2) a TR locus overlaps in position with a TE by at least one

period. To identify TRs that are at least 70% similar to a TE, the targeted TR repeat sequences were aligned against the TE consensus database using BLAST by setting e-value at 10^{-6}, mismatch penalty at -1 and word size at 7. In the second method of identification, the starting and ending genomic positions of a tandem repeat arrays were cross-checked using an in-house PERL script. Any TR overlapping a TE by the length of at least one TR period was considered TE derived. Clustering all selected TRs was performed by using the NCBI BlastClust tool with a maximal sequence length disparity of 10% and a minimal sequence similarity of 85% among the members of a cluster.

3.2.3 Identification and distribution of TE families contributing to TR

The TR repeat unit was aligned pairwise with its corresponding candidate parent TE using the NCBI bl2seq tool with zero penalty for alignment gap to identify the region of the TE that is duplicated. The contribution of each TE family and subfamily to TR is evaluated not only by the total number of TRs contributed, but also based on the relative TE abundance, which is represented as the percentage of TE in the subfamily that are contributing to TR. This relative number is calculated by dividing the actual number of TE loci involving TR with the total loci of that TE and multiplying by 100.

3.2.4 Identification of sequence similarity among repeat units and with orthologous sequences in other primate genomes

To identify the possible mechanism of TR expansion, 5 AluJ-derived TRs with more than 15 repeat units were randomly chosen for manual analysis. Each individual repeat unit was aligned to hg19 using BLAT with default parameters to identify all genomic regions that it matches with. All aligned regions were sorted according to the similarity score to identify the best match. If the expansion occurred due to sequential duplication

of the repeat unit, the best matching region would be the repeat unit adjacent to the test sequence. If a TR was generated along with retrotransposition, i.e. simply representing a copy of a TR in the parent TE somewhere else, then we would expect to see better sequence similarity elsewhere in the genome than among repeats in the same array. The tandem arrays were then aligned with the latest version of chimpanzee, orangutan, gorilla and marmoset genome sequences using UCSC genome browser in an attempt to find similar repeat arrays in other primates. If the expansion occurred slowly through evolution, each repeat array was expected to have partial to no match with other primate genomes. Moreover, TRs with higher number of repeat units was expected to had accumulated more mutations than TRs with smaller number of repeat units due to their residence in the genome for a longer time. To test whether TRs with a larger number of repeats are older than the TRs with a small number of repeats, we surveyed the maximum sequence divergence among the repeat units in TRs. To do this, we classified all non-LTR12 and non-L1PA TE-derived TRs in two classes: one with ≤3 units and the other with ≥10 units. Repeat units in each TR were then separated using Perl script and aligned pairwise to one another to create an evolutionary distance matrix among the repeat units using CLUSTALW (downloaded for Linux platform from ftp://ftp.ebi.ac.uk/pub/software/clustalw2) (Chenna *et al.*, 2003). The distance is calculated by dividing the total number of mismatches between two units with total number of matched pairs. The maximum divergence for each TR was obtained from its corresponding distance matrix.

3.3 Results and discussion

In this study, we seek to perform a genome-wide survey of the contribution of transposable elements to the generation of tandem repeats and examine the possible mechanisms. The starting point of this study consisted of the output data from the Tandem Repeats Database which provides a compilation of all tandem repeats in the human genome ranging from 1bp to 2000 bp in size of the repeat unit. For the latest assembly of human reference genome (NCBI build 37 or Hg19), TRDB annotates 31,472 minisatellites and satellites (both will be called minisatellites hereafter for simplicity) with repeat unit length more than 20bp, minimum GC content of 40%, minimal number of repeats of 2 and have at least 95% identity among the repeat units in an array. A minimal 40% of GC content was applied to eliminate TRs that contain mainly low complexity or simple repeat sequences, which can derive from poly (dT) or poly (dA), present frequently in non-LTR retrotransposable elements as the 3'-end polyA track or the internal sequence of Alu or SVA. Of the 31,472 minisatellites, 7,276 (23.12%) were detected as being derived from transposable elements either by sequence similarity with TE consensus sequences or by overlapping an annotated genomic TE region by at least one period. The TE-derived minisatellites were then classified into 5,932 clusters based on their sequence similarity, with each cluster representing tandem repeats that are likely to have derived from or related to a particular TE. Among the 5,932 clusters, 185 contain similar sets of tandem repeats that are found in more than one locus in the whole genome and thus are termed as multi-locus TRs or "mlTRs" following the nomenclature proposed by Ames et al. (Ames *et al.*, 2008), and 5,747 clusters contain TE derived TRs that are present only in one locus in the genome and thus are termed as single-locus TRs or

"slTRs". These 7,276 TE-derived TRs contribute to a total of 1.05 Mb of sequence or ~0.32% of the human genome, and we believe that these numbers represent a underestimate of such events that have happened in the human genome, since we may fail to detect a lot of old TRs as a result of high sequence divergence (see more discussion later).

3.3.1 Younger and more active TEs are more susceptible to tandem duplication

Almost 19% of the TE-TRs (1,374 of 7,276) are derived from LTR12 and L1PA subfamilies of retrotransposons. This was expected due to the internal tandem repeat in the consensus sequence of these two subfamilies. To avoid bias in assessing the general trend, we treated these separately from those associated with other TE subfamilies. For the other TEs, the most number of TRs (2663) were found to be derived from Alu, while ERVs and L1 had 1597 and 601 associated TRs, respectively. Since the abundance for each TE subfamily is different in the human genome, the number of TEs for each subfamily of TEs was normalized for the total number of TEs in that subfamily in the genome. After normalization, Human Endogenous Retroviruses (HERVs), including the internal viral sequences and LTRs, exhibit a relatively higher percentage of tandem duplication (39%), with almost 90% of members belonging to HERV-K subfamily, which is the youngest and most active ERV. Even though the actual number of SVA-derived TRs is as small as 12, when normalized, SVA has the second highest relative abundance (32%) in terms of generating TRs. Following HERV and SVAs, Alus are the TE classes with the third most abundant tandem repeats, and all of them belong to the younger and more active classes of TE in the human genome (Figure 3.1a). When the subfamilies of Alu are examined for relative abundance of tandem repeats, all subfamilies

exhibit somewhat similar abundance, with AluY seeming to show slightly higher abundance (Figure 3.1b). However, the mean abundance of the three major subfamilies of Alu – AluJ, AluS and AluY – shows a clear increment of relative TR abundance from AluJ (0.18) to the intermediate AluS (0.24) to AluY (0.40). This also follows the trend of younger/more active TEs generating a higher number of TRs as AluJ is the oldest subfamily of Alus, while AluY is the youngest and most active subfamily of Alus. The age of AluJ has been dated back to 26 million years ago (Kapitonov & Jurka, 1996) and no species-specific AluJ activity has been identified in the comparative studies between humans and chimpanzees. AluS diverged from AluJ later and only 262 new AluS insertions have been identified in humans that happened within last 6 million years ago, which is a fraction of the total AluS insertions annonated in the human genome (Mills *et al.*, 2006). The youngest family of Alus is AluY and they are believed to be the most active Alu family in the present human genome. The trend of increasing relative TR abundance from older subfamilies to newer subfamilies of TEs may indicate that the initiation of TE-derived TRs, at least for a large number of cases, can potentially be associated with the retrotransposition process of TEs. In other words, the positive association between abundance of TE-derived TRs and transposition activity level of TEs may suggest that retrotransposition contributes to the initiation of TRs, despite the possibility that the lower relative abundance of TRs on older TEs could also be due to recombination-mediated deletion and/or lower detection because of sequence divergence.

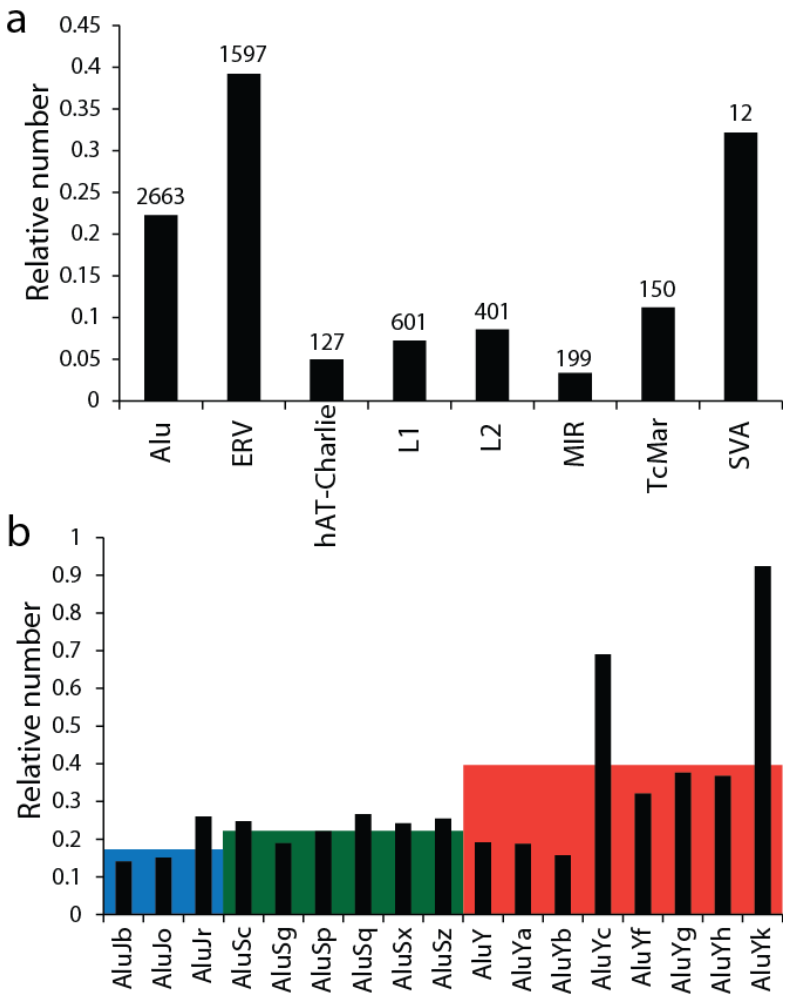


Figure 3.1 Relative abundance of major families and subfamilies of TEs that generate TRs. Relative abundance is calculated by dividing the number of TE-derived minisatellites by the total number of members in that TE family. Panel a: Relative abundance of major families of TR-associated TEs. The actual number of TE-derived TRs is at the top of each bar. Panel b: Relative abundance of Subfamilies of TR-associated Alus. The color shaded boxes are average relative abundance for the group with blue for AluJ, green for AluS, and orange for AluY. It is evident that the average relative abundance increases from AluJ to AluS to AluY.

3.3.2 Older TEs have a larger number of repeat units than younger ones

The initiation of TR expansion occurs more often with younger classes of TEs (Figure 3.1). However, once a region is repeated at least once, the increase in the number of the repeat may occur by previously reported mechanisms for such events (further

discussed later in the section). When the number of repeats for each major subclass of Alu is plotted in a graph, a steady decrease in number of repeats from older to newer class of Alus become clear (Figure 3.2). The AluJ has a mean number of repeat units of 2.42, AluS has 2.31 and AluY has 2.30. There were differences in variance among these classes of Alus ($P<0.0001$). However, there were no differences identified in mean number of repeat units between AluS and AluY in a two-tailed t-test. But this can be largely due to the fact that the total number of TRs generated by AluS is more than four times higher than that by AluY with majority having a repeat number below 3. Furthermore, the evolutionary distance between AluS and AluY is less than that between AluJ and AluS (Churakov *et al.*, 2010). When older AluS subfamilies (AluSx, AluSg, AluSp and AluSq) were examined, 8.11% of their associated TRs have more than 3 repeat units, while only 6.70% of TRs from AluY have more than 3 repeat units (data not shown) and the newest AluY elements – AluYa and AluYb have no TRs with more than 3 repeat units. This decrease in repeat number from older to younger families of TEs can be explained as the expansion of repeat units is a slow process, and it takes longer time to generate more TR repeats.

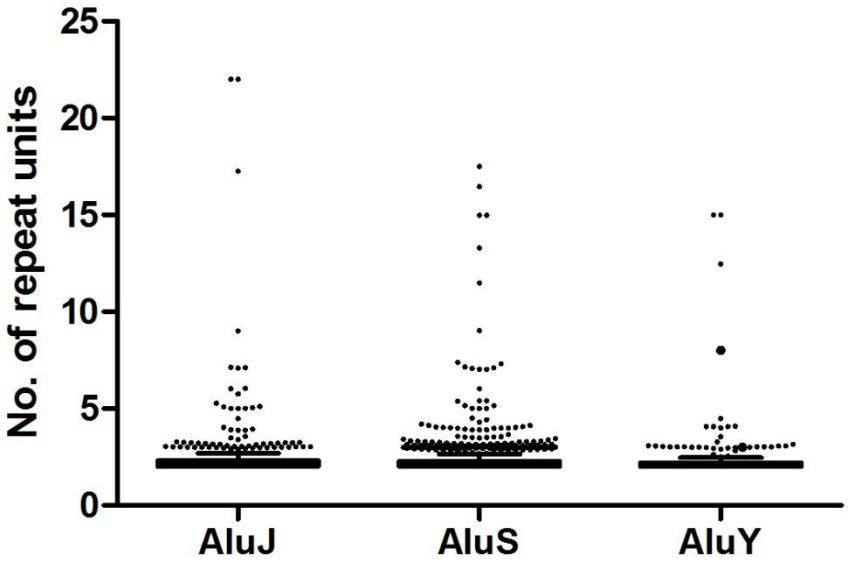


Figure 3.2 Box and Whiskers plot of the number of repeats for TRs derived from the three major classes of Alu. The average number of repeat units decreases from AluJ (2.42) to AluS (2.31) to AluY (2.30).

When the TE-derived TRs with a larger number of repeats were aligned against the orthologous sequences from other primates, only a portion of the total repeat is found in the outgroups. In Figure 3.3, a 17 tandem repeats of 52 bp from AluJo (from 226 to 278 bp of the consensus sequence) is aligned against the corresponding sequences in the outgroup genomes, and only a portion of the total TR are matched in these genomes. Since AluJo appeared in primates 26 million years ago (Kapitonov & Jurka, 1996), the extra repeat units can be explained as further extension of the common repeat units in the human genome after the diversion from chimps by *in situ* duplication rather than by transposition. This is further supported by our observation in examining 5 randomly chosen Alu-derived TRs with a minimal number of repeat units of 15, in which the repeat

units in an array of TR are best aligned against each other than any other region in the genome, indicating that one unit was used as the source of the other for duplication in a local manner. When the mlTRs were investigated, 45 out of 185 mlTRs were found to be variable in number of tandem repeat units in different loci. With exception of one, all of these mlTR clusters follow the same trend of decreasing number of loci with increase in the number of repeat units (Table 3.1). This again indicates that the expansion of repeat units of a TR may occur sequentially with time, for which in a cluster of mlTRs, the TRs with higher number of repeat units are seen in lesser number of loci.

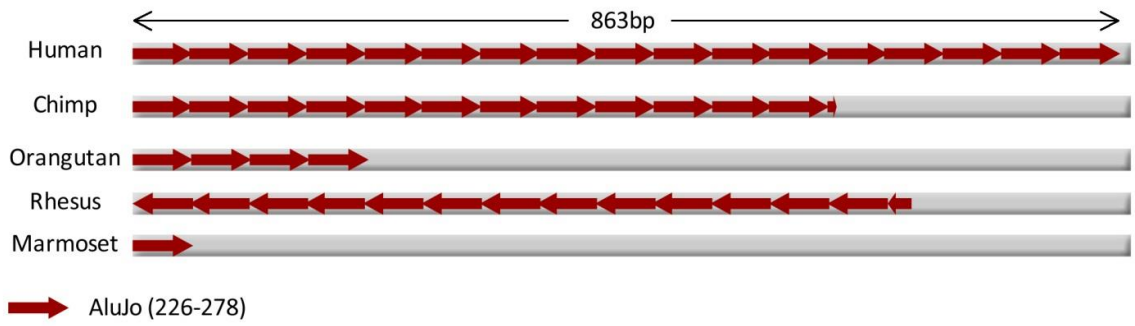


Figure 3.3 A schematic comparison for a 17-repeat TR array involving the 226-278bp region in a AluJo among difference species. The human genomic region was compared with the corresponding region from chimp, orangutan, rhesus and marmoset.

Table 3.1 The number of mlTRs at different repeat units for mlTR clusters

Cluster No.	Unit size	No. of repeat unit of each TR in the cluster	No. of different loci the TR appear in	TE subfamiliy
6	109	2	36	AluS
		3	6	
		4	1	
		17	1	
7	125	2	36	AluS
		3	5	
		5	3	
20	22	2	22	L1P3/L1PR
		3	1	
14	268	2	25	AluY
		3	3	
22	32	2	13	AluS
		3	1	
		7	1	
18	161	2	21	AluY
		3	1	
		14	2	
24	54	2	11	THE1C-int
		3	1	
32	68	2	8	HERVH-int
		3	1	
71	352	3	2	AluSx
		4	2	
55	42	2	4	HERVH-int
		3	1	
		6	1	
84	49	4	3	MER57A-int
		17	1	
259	42	2	1	HERVH-int
		4	1	
89	28	2	3	LTR10F
		8	1	
31	113	2	8	Harlequin-int
		3	1	
208	71	2	1	AluSx
		3	1	
62	42	2	3	HERVH-int
		3	1	
		4	1	
67	31	2	4	AluY/AluS
		3	1	
198	78	2	1	AluS

		3	1	
139	664	2	1	HERVE-int
		3	1	
129	24	2	2	HERVH-int
		3	1	
236	45	2	1	AluSx
		3	1	
145	255	2	1	MSR1
		4	1	
178	102	2	1	AluS
		3	1	
124	28	2	2	LTR10F
		12	1	
47	114	2	4	HERVH-int
		3	1	
		5	1	
43	25	5	1	L1M5
		6	1	
244	44	2	1	AluSq
		3	1	
74	221	2	3	AluY
		3	1	
113	39	2	1	AluY/AluS
		3	1	
39	64	2	1	HERVH-int
		4	2	
		5	1	
		6	1	
		7	2	
120	32	2	2	AluS
		3	1	
239	49	3	1	L2c
		5	1	
301	32	13	1	LTR7B
		14	1	
27	32	2	10	AluS
		3	1	
60	63	2	4	HERVH-int
		3	1	
247	46	2	1	AluSx
		4	1	
82	49	2	3	AluS
		3	1	
212	70	2	1	AluJr
		3	1	
33	61	2	5	HERVH-int
		3	4	

63	37	2	3	HERVH-int
		3	1	
		6	1	
80	93	2	3	HERVH-int
		3	1	
54	42	2	2	HERVH-int
		3	3	
		8	1	
40	60	2	5	HERVH-int
		3	1	
		9	1	
278	36	6	1	MSR1
		29	1	
34	30	2	7	AluS
		3	2	

When LTR12-derived TRs are analyzed, the number of repeats in the internal sequence is found to be variable throughout the genome. Complying with the relationship seen between the number of repeats and number of occurrence in non-LTR12 mlTRs, the larger the number of repeated sequences, the less the number of loci. This provides evidence that these duplication events have taken place throughout the evolution and the repeats are possibly increased sequentially in number. Also for this reason, an entire TR generated by the older TEs or part of a TR that has existed for much longer time have been subject to more mutations/deletions than the younger ones. In other words, the TRs with more repeat units should accumulate more mutations than TRs with smaller number of repeat units because of their longer residence in the genome. When the evolutionary distance among repeat units in TRs with ≤3 repeat units and ≥10 repeat units was examined, the mean highest distance found in TRs with ≤3 units was 0.5330 while that of TRs with ≥10 units was 0.8049 (Figure 3.4). The difference in maximum divergences among repeat units between the short and long TRs is statistically significant (two tailed t-test $P<0.0001$). This provides direct evidence that TE-derived TRs are expanded

gradually throughout evolution. Some of these TRs or TR repeats may have been mutated to a point where they have become undetectable as tandem repeats by the current algorithms. For this reason, the number and/or the length of TRs derived from TEs may have been underestimated.

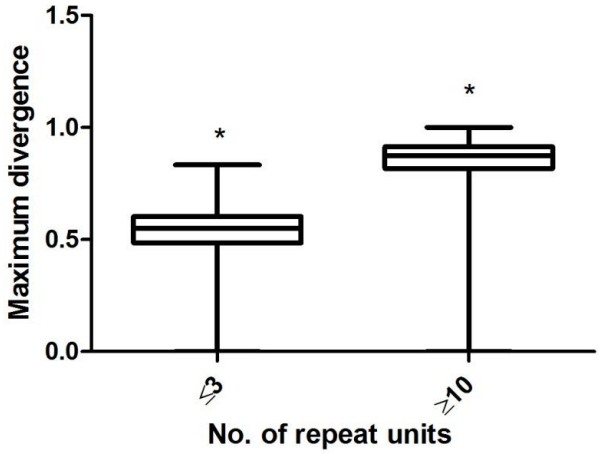


Figure 3.4 Box and Whiskers plot of maximum divergence among repeat units in TRs with ≤3 and ≥10 repeat units. The mean maximal divergence among repeat units in TRs with ≤3 units is 0.5330 and is 0.8049 in TRs with ≥10 units based on all 5,902 non-LTR12 and non-L1PA TE-derived TRs. The asterisk denotes that they are significantly different ($P<0.0001$) in a two-tailed t-test.

3.3.3 Certain TE regions can act as hotspots for tandem duplication

To see whether hot spots of TRs exist in the genome or in specific region of TEs, we plotted the TE-derived TRs in the whole genome, and no obvious hotspots were seen in the genome (Figure 3.5). When the positions of the repeated regions are plotted in AluJ and AluY, no TR hotspot was identified (Figure 3.6a,b). But there are two regions (59 to 137bp and 176 to 206bp) found in the AluS consensus sequence that are spanned by comparatively more TRs than other regions (Figure 3.6b). There are also two distinct hotspots observed for LTR12 from 99 to 182bp and from 719 to 841bp (Figure 3.6c).

This may be due to the fact that TR existed in the original LTR12 sequences and the TR were propagated also by transposition, different from other TE-derived TRs where initiation and expansion occurred at or after individual TE insertion.

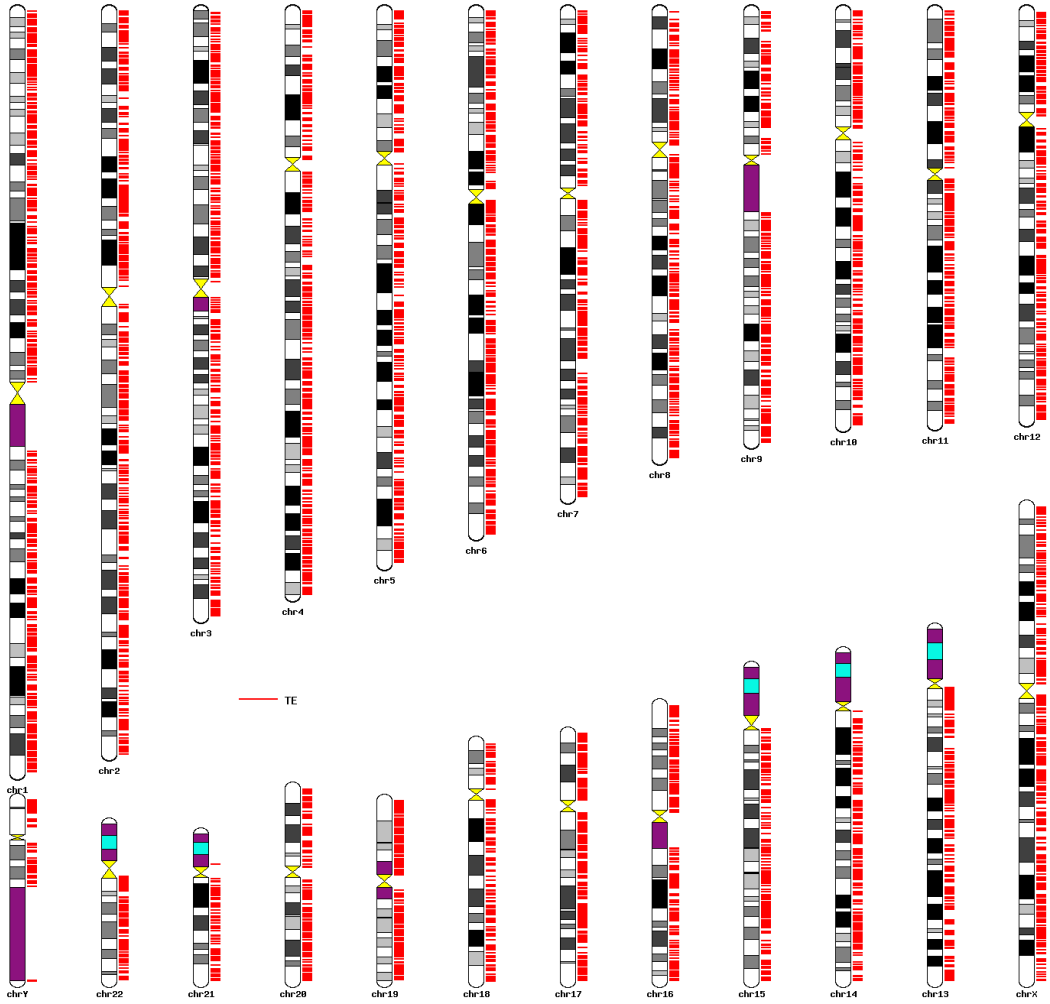


Figure 3.5 Genomic locations of all TE-derived TRs. All individual TRs are plotted onto the human chromosome ideogram based on their genomic positions in the UCSC hg19 reference sequence. Chromosomal regions in color are heterochromatin regions which mostly lack sequence. Despite the ubiquitous but non-homogeneous distribution of TE-derived TRs in the genome, there seem to have no obvious hot spot for TR generation in any part of the genome.

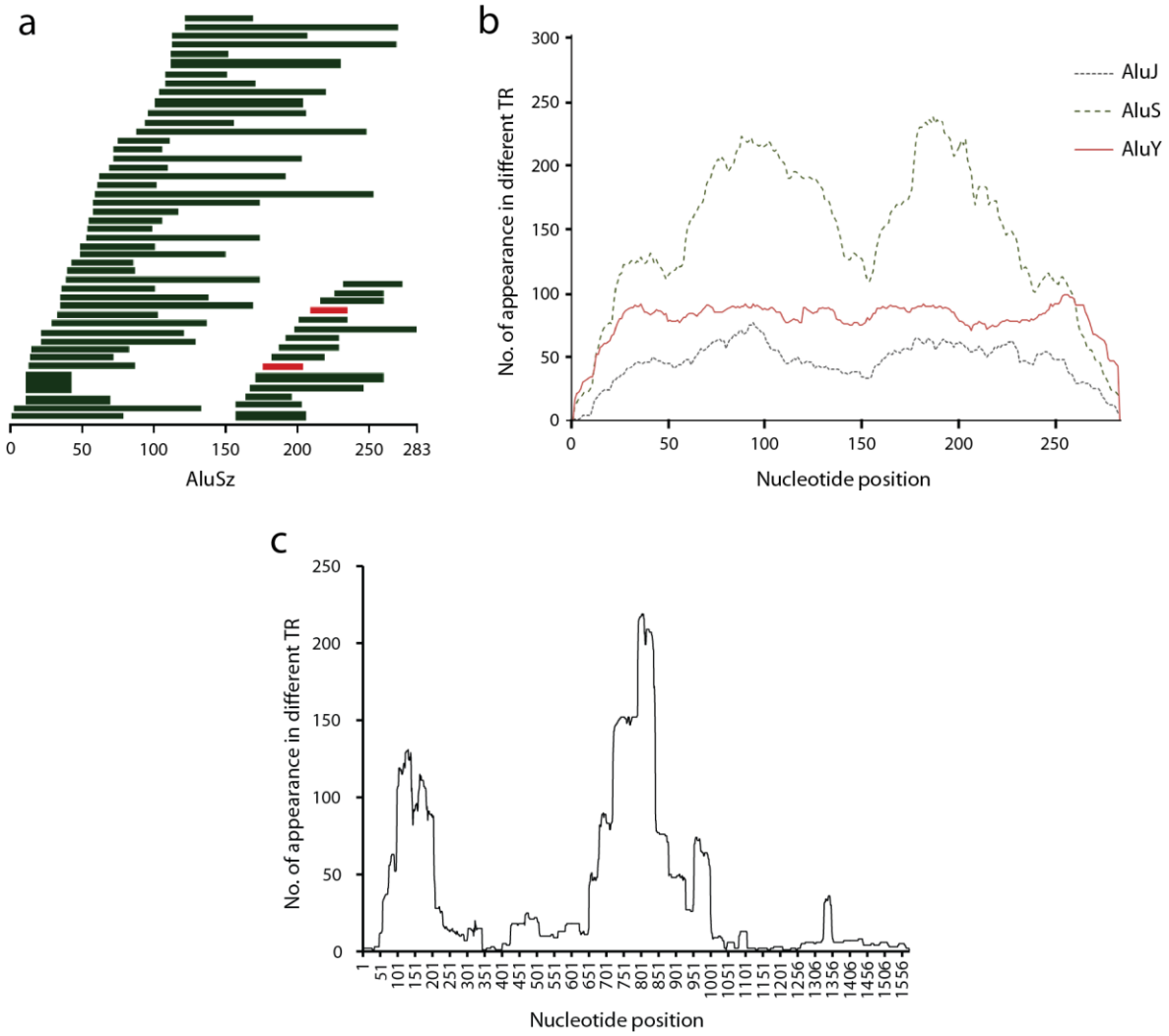


Figure 3.6 Regions of TE that are involved in generating TRs for Alus and LTR12. Panel a: Representation of a selected number of fragments of AluSz that have generated TRs. Selection was made randomly to demonstrate that the repeat can occur from any region of a TE. The height of each bar is proportional to the number of repeats. Green colored regions are duplicated in 2 loci and red colored regions are duplicated in 3 loci; Panel b: The number of TRs spanning each nucleotide of AluS, AluJ and AluY; Panel c: The number of TRs spanning each nucleotide of LTR12.

3.3.4 Multiple mechanisms for generation of TE-derived TRs

Of the 7,276 TE-derived TRs, the terminal repeat unit in 159 TRs is incomplete. These truncated terminal units are smaller in size than the other unit(s) in the same TR by maximum of 10%, i.e., if the unit length of the TR is 100bp, the terminal unit's length is between 90 to 99bp. Initiation of these TRs can follow the mechanism of slipped-strand mispairing proposed by Taylor et al. (Taylor & Breden, 2000), as having an incomplete or truncated repeat unit at the end of the repeat array is a key feature of that mechanism. Among other TE-derived TRs, 300 were found to have flanked by direct repeats of size 5-20 bp. The initiation of such TRs can be explained by the mechanism proposed by Haber and Louis (Haber & Louis, 1998). According to that model, replication slippage including gene conversion or unequal crossing over during meiotic replication can cause gain or loss of a copy of the region flanked by such small direct repeats. The majority of these flanking repeats is of size at 7 bp, which is consistent with this model (Table 3.2) (Jurka & Gentles, 2006; Nishizawa *et al.*, 2000). These two established mechanisms may explain initiation of only 6% of all TE-derived TRs. The rest 6,817 of the total of 7,276 TE-derived TRs are not flanked by direct repeats or incomplete terminal repeat, with the majority have only two repeat units. Thus these 6,817 TRs are unaccountable by the currently established mechanisms, and hence are likely subjected to one or more yet to be identified mechanism(s). Among these, 136 TRs exhibit a specific pattern of repeat of a partial Alu (average length of 88.6 bp) adjacent to a full or near full length Alu (at least 300 bp). The duplication of the partial Alu sequence at the 5' end of a TE may occur due to recombination or unequal crossing-over due to the presence of an endo-nucleolytic site immediately adjacent to the 5' end of the TE. This endonucleolytic site is the target of LINE-1 endonuclease and can function as recombination hotspots (Babcock *et al.*, 2003).

It has also been proposed that when the endonuclease acts on such targets, single-strand nicks can be generated in DNA to promote recombination (Gentles *et al.*, 2005). In addition to such well-defined pre-integration endonuclease target sequences, potentially kinkable dinucleotides such as TA, CA and TG can also promote nicking, consequently promoting recombination (Jurka *et al.*, 1998; Mashkova *et al.*, 2001), and thus may serve as potential mechanism of TR initiation.

Table 3.2 The distribution of direct repeat length for TE-derived TRs with identifiable direct repeats.

Direct repeat length (bp)	Number of occurrence
<7	0
7	145
8	47
9	21
10	11
11	12
12	15
13	12
14	9
15-20	28

3.4 Conclusions

While transposable elements are known for genomic rearrangement and expansion of the genome by transposition, we show in this study that they also play a role in genome expansion and alternation by contributing to tandem repeats. Over 20% of all minisatellites/satellites are contributed by TEs, constituting a total length of 1.05 million base pairs in the human genome, and according to the results of this study, this number is and will be increasing.

Results from this study suggest that the tandem repetition of full or partial TEs can be triggered during retrotransposition and once it is duplicated, the expansion of the repeat units can slowly occur through time. While a small portion (6%) of TE-derived TRs can be explained by one of the mechanisms postulated so far, the mechanism(s) for the majority is yet to be identified, thus our results present the need for identifying new mechanisms underlying the TE-derived TRs initiation and expansion. Furthermore, no study has yet revealed the detailed nature of the recombination hotspots adjacent to the minisatellites in terms of their DNA primary structure, plasticity or secondary structure, thermal stability, or functionality (Murray *et al.*, 1999). Understanding these phenomena will definitely help identifying exact mechanism(s) of tandem repeats derived from transposable elements.

Chapter 4 : Construction of a genome sequence ancestral to all modern humans

(Part of this chapter is reprinted from the manuscript Ahmed M, Liang P: Construction of a genome sequence ancestral to all modern humans. Manuscript in revision.)

4.1 Background

Genome sequence among individual humans varies in multifarious aspects. The release of the complete whole genome sequences (Lander *et al.*, 2001; Venter *et al.*, 2001) provided, for the first time, a panoramic picture of the human genome, which has since aided the progress of a plethora of genome-wide high throughput technologies for genomic and functional analyses. The reference genome represents a single or haploid version of genome sequence, which in turn represents a consensus genome sequence of a few individuals, mostly of a Caucasian background. Therefore, the reference genome sequence itself provides little reflection of the variations in genomic sequences that occur naturally between individuals or populations. Rather, it can be used as a reference for identifying variations via comparative genomics. The recent advent of new genomic sequencing technologies, particularly the Next Generation Sequencing (NGS), has allowed sequencing of more and more individual genomes that subsequently has made the identification of various sequence variations among individual genomes feasible (Metzker, 2010). The human genome project (Lander *et al.*, 2001), the SNP consortium (Sachidanandam *et al.*, 2001), the International HapMap project (International HapMap Consortium, 2005), and 1000 Genome Project (1000 Genomes Project Consortium *et al.*, 2010) have collectively identified more than 15 million sequence variations in the human genome, the majority of which are contributed by Single Nucleotide Polymorphisms (SNPs). The 1000 Genome Project (1KGP), launched in 2008, has been the most recent and largest project to date to have conducted comparative sequence analysis among individuals from all major populations – West Africa, Europe, East and South Asia, and the America. The pilot phase of the project alone has cataloged ~8.4 million novel SNPs

and ~840,000 novel insertions/deletions (1000 Genomes Project Consortium *et al.*, 2010). The results from all of these large scale comparative studies have enabled us to realistically picture the existing level of sequence variations in our species both at inter-individual and inter-population level, most often with information of the exact genomic location.

Along with SNPs, Structural Variations (SVs) have been revealed as another major type of variation observed in humans that include Copy Number Variation (CNV), insertions/deletions, and tandem duplications, translocations, and inversions, as well as transposable element insertions (TEIs). The current methods of identifying sequence variations involve aligning the NGS output sequences for an individual genome with the reference genome and seeking anomalies in alignment. For analysis of SVs, this method only provides information about structural variation relative to the reference genome, i.e., an "insertion" is the presence of an extra block of sequence in the individual genome compared to the reference genome, and a "deletion" is the absence of a block of sequence in the individual genome compared to the reference genome. Thus, an insertion/deletion or copy number gain/loss does not necessarily mean a true insertion/deletion or copy number gain/loss by the order of events during human genome evolution, and this has led to a lot of confusion in defining these terms and difficulties in making sense of these SVs. In this study, we propose a minimalistic version of a genome sequence that represents the most recent Common Ancestor to all modern Human Populations (CAHP) by utilizing all human sequence variation data available to date and genome sequences of chimp, gorilla and orangutan (Figure 4.1). The genome sequence of CAHP can serve as a better reference genome sequence in comparative genomics and evolutionary studies by

providing an accurate definition about the type of any genomic variants with respect to the ancestral state of the variant. This genome sequence can also be useful in estimating the minimal distance and genetic differences between human and other primates, as well as with archaic humans, e.g., Neanderthals and others, as their genome sequences become available. Our study indicates that a total of at least 8.89 million bases of DNA have been inserted in the most recent major release of the human reference genome (assembly version ID GCh37) since the most recent common ancestor to all modern humans. While large insertions contribute the most to the size increase (approximately 68%), mobile elements are the most abundant type of insertion (at 2,071 loci) and over 320,000 single nucleotides in the reference genome are different than in the genome sequence of CAHP. These data should shed light on the major genetic events involved during the evolution of modern humans.

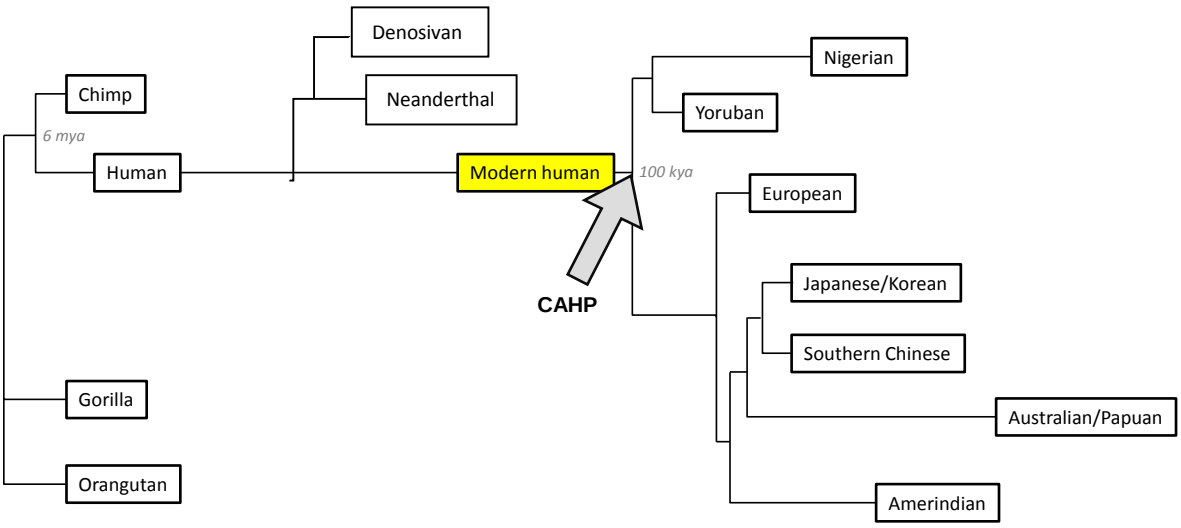


Figure 4.1 A schematic representation of the evolution of modern human populations. The branches are not drawn to scale. The phylogenetic relationships are drawn based on information presented in (Bowcock *et al.*, 1994; Nei & Roychoudhury, 1993; Ohshima *et al.*, 2003) and references therein. CAHP, Common Ancestor of all modern Human Populations.

4.2 Materials and Methods

4.2.1 Construction of the genome sequence of the most recent CAHP

The foundation for the construction of the genome sequence of the most recent CAHP is based on identifying the regions in the human reference genome that are variable among different individuals and then using other primate genome to determine the ancestral state of each of the variable regions. Any change in the human reference genome since the most recent CAHP should not be fixed in all individuals and thus should be polymorphic in at least two individuals from different families. The construction of the ancestral genome is based on removing the sequences from the reference genome that are inserted or adding the sequences to the reference genome that were deleted after divergence of all modern human populations. In this study, we define an **INSERTION** in reference to the CAHP as a block of DNA sequences that is present in the reference genome or another personal genome, but are absent in the genomes of multiple human individuals AND also absent in the genomes of chimp and other primates (Table 4.1). This is equivalent to a "deletion" in the test genome that misses the sequence based on the current reference genome. A **DELETION** in relation to the CAHP is defined as a block of DNA sequences that is absent in the reference genome or another personal genome, but present in multiple human individuals and at least one non-human primate genome. This is equivalent to an "insertion" in the genome that carries the sequence based on the current reference genome.

Table 4.1 State of the sequence in chimp, individual human genomes, reference genome in the event of insertion or deletion in relationship to CAHP.

	INSERTION IN THE REFERENCE GENOME	DELETION IN THE REFERENCE GENOME
Chimp	Sequence absent	Sequence present
Human 1	Sequence absent	Sequence present
Human 2	Sequence absent	Sequence present
Human Reference Genome	Sequence present	Sequence absent
CAHP	Sequence absent	Sequence present

We identified these structural variations in the reference genome compared to at least two individual test genomes and chimp genome, and the method of detection of each type of variations is described in subsequent sections.

4.2.1.1 Identification of Large Insertions

The large insertions were obtained primarily from data produced by the Structural Variation team from the 1000 Genome Project (Mills *et al.*, 2011). The original data were published for reference genome assembly Hg18, and they were converted for the latest assembly (Hg19) by using LiftOver (Hickey *et al.*, 2013). The redundant entries and insertions observed only in one individual were removed using in-house Perl scripts. To confirm the ancestral state, 100 bp flanking sequences from both sides of each insertion were mapped against the chimpanzee reference genome assembly panTro3 using BLAT (Kent, 2002) requiring a minimal identity of 90% and a minimal sequence coverage of 90%. Some insertions reported in the 1KGP do not have the exact breakpoint information in the genome, rather reported with a sequence position range for each of the two breakpoints. For these insertions, 100 bp flanking from both sides of the outer boundaries were mapped against the chimp genome sequence. If the gap between the mapped

flanking is not larger than the combined length of the sequence range for both breakpoints, the mapped regions including the gap sequence from the chimp reference genome is aligned back to Hg19 using BLAT. The insertion is confirmed if the mapped sequence in Hg19 is split into two parts with a gap in between. If the positions of the start and end of the gap lie between the two breakpoint ranges originally reported, that confirms exact breakpoints for that particular insertion. All of the steps are summarized in Appendix II Figure 1. The final list is cross-checked against other kinds of insertions identified (described in the following subsections) to avoid redundancy, and cross-checked also with the positions of all pseudogenes in the human genome using an in-house Perl script. Among all large insertions identified in the reference genome compared to the most recent CAHP, 21 insertions are found to be pseudogenes. Additional 226 such insertions are found to overlap with TE insertions (described later), thus they were omitted from the final list. The insertions for which the exact breakpoints could not be ascertained were stored in a separate list for future reference.

4.2.1.2 Identification of Large Insertions

The primary candidates are obtained from the 1KGP data (Mills *et al.*, 2010) with insertions observed only in one individual removed. 100 bp sequences from both sides of each insertion were mapped to the chimpanzee reference genome. The gap between successfully mapped flanking sequences in chimp for each insertion was compared with predicted insertion size in the test human genome, and events in which the gap size in chimpanzee genome between the mapped flanking sequences is within 90% of the length of predicted insertion size are considered as candidate deletions in the reference genome compared with chimp. Each of these insertions is then manually checked for presence in

the orangutan and rhesus reference genome sequences using the UCSC genome browser. For the final candidates, each insertion sequence obtained from the chimp genome along with 100bp flanking sequences was mapped back to human reference genome to identify the exact position of the insertion and to obtain the inserted sequence. These sequences are present in other primates and in at least two individuals but absent from the human reference genome, thus they are likely deleted in the reference genome since the CAHP.

4.2.1.3 Identification of TE insertions

A similar approach was taken for identifying TE insertions in the reference genome compared to CAHP. The list of TE deletions observed in individual genome sequences compared to the reference genome was obtained from 1000 Genome Project data (Stewart *et al.,* 2011) and dbRIP (Wang *et al.,* 2006b). Data from the two sources were treated separately to detect candidate TE insertions in the reference genome compared to the genome sequence of CAHP. The final list of TE insertions was obtained after removing 732 redundant entries from the two sources and then by combining overlapping insertions into larger contigs.

4.2.1.4 Identification of TE deletions

An approach similar to detecting large deletions is taken for detecting TE deletions. The initial list of TEs that are found present in at least two individual genomes but absent in the reference genome was obtained from the 1KGP data (Stewart *et al.,* 2011). A 100bp sequence from each end of such deleted TEs was obtained from the reference genome and aligned to chimp and gorilla genomes. The size of the gap between mapped flanking sequences for each deletion was compared with the length of deletion originally reported. If the sizes are similar and if the chimp and/or gorilla genome contains a TE in

the orthologous position, which is from the same TE subfamily reported in the human, it is likely that the particular TE got deleted in the reference genome since the most recent CAHP. The deletion in the reference genome is further confirmed by manually detecting the pre-integration site requiring the presence of one copy of the TSDs in the reference genome.

4.2.1.5 Identification of small insertions

The small insertions are obtained from dbSNP (Sherry *et al.*, 2001) ranging from 1bp to over 200bp that are polymorphic for presence or absence in human individuals. Sequence information in the orthologous position in the chimpanzee genome was obtained from the UCSC Genome Browser (http://genome.ucsc.edu), which used the UCSC tool LiftOver (Hickey *et al.*, 2013) to generate such data. The sequences of the small insertions in the reference genome were compared with the orthologous sequence information in the chimpanzee genome in the LiftOver output file using a custom Perl script. Any insertion that is not present in the orthologous position in chimpanzee is considered as an insertion event that took place in the human lineage since the CAHP. These inserted sequences are removed from the reference genome to construct the genome sequence of the CAHP. Many small insertions that were found to have been inserted in the reference genome since the CAHP are overlapping or positioned in tandem order in the reference genome. These overlapping or tandem small insertions are joined together to form contigs by using an in-house Perl script. Each contig is given a unique ID. The dbSNP IDs that each contig contains are saved in a separate file for referencing back.

4.2.1.6 Identification of tandem repeat expansion

The tandem repeat information for the human reference genome is obtained from the Tandem Repeat Database (Gelfand *et al.*, 2007). The tandem repeats are filtered by three criteria: 1) a total length of 30bp or more; 2) a minimum of two repeats; and 3) a sequence similarity score of at least 98% among the repeating units. A 100bp of flanking sequence on both side of each TR locus was obtained from the reference genome and mapped against the chimpanzee genome sequence using BLAT. TRs with at least one repeating unit missing (i.e., TRs with at least one extra repeating unit in humans) are kept for further analysis. These extra blocks of repeating units could be results of either expansion in the human lineage, or deletion in the chimpanzee lineage. To identify the former events and to confirm that the expansion occurred only after the divergence of CAHP into different populations, candidate TRs are compared with large insertion data previously identified to be inserted in the reference genome since the CAHP.

4.2.1.7 Determination of the ancestral state of SNVs

The ancestral state of all single nucleotide variants is determined by sequence information in the orthologous position in chimp, gorilla and macaque. Such information is obtained from a custom track from UCSC genome browser that lists all SNVs with orthologous variant in those three outer primates determined by LiftOver. Ancestral nucleotide is determined as the one, which is most common among the three other primate genomes. If the nucleotides at the orthologous positions of all three primates are not similar to one another, preference for determining the ancestral nucleotide is decided in descending order of chimpanzee, orangutan and macaque. This data includes only

biallelic single nucleotide polymorphisms that can be mapped to only one locus in the human genome and are not associated with any clinical conditions.

4.2.1.8 Combining overlapping entries from various databases

After identifying all small insertions from dbSNP, all transposable element insertions from dbRIP and 1KGP, and all large insertions from 1KGP that are absent in chimpanzee, many such insertions are found to overlap with one another. We combined the overlapping entries into one entry and given a custom ID to each combined insertion. We combined 160, 519 overlapping SNP/indel entries into 69,547 combined IDs, 73 overlapping 1KGP large insertion entries into 22 combined IDs and 5 overlapping TE insertions into 2 combined IDs. Among 69,547 combined SNP/indels into contigs, 242 contigs have a length over 30bp, and these were moved to the dataset for large insertions.

4.2.1.9 Assembling the final genome sequence

The final lists of all variations are cross-checked and combined for positional overlaps to avoid redundancy. Each type of variations are saved in a separate file in gvf format. The SNVs are then replaced with their ancestral state in the reference genome, the insertions are removed from and deletions are added to the reference genome sequence by an in-house Perl pipeline. The Perl pipeline is coded in reusable and customizable way to generate the whole genome sequence into individual chromosome from a set of gvf files. This pipeline is also useful for updating the genome sequence of CAHP as more variant data are available. The gvf files are also converted to bed formats using an in-house script, which can then be used to generate a custom track for UCSC browser for visualization of the changes, as well as to visualize the ancestral state of any given genome location specified by the users.

4.2.2 Analyses of the genome sequence changes from CAHP to the current human reference genome

Size distribution of all insertions was analyzed using custom Perl scripts. The flanking sequence analysis was performed using "DR Finder", a robust algorithm developed for the project, and it identifies direct repeats flanking any given genomic region. The pipeline is discussed more in the subsection 4.2.4. The mechanisms for large insertions are predicted by the tool breakseq (Lam *et al.*, 2010). The population information for large insertions and the allele frequency of TEs in different population were retrieved from the 1000 Genome Project data (Stewart *et al.*, 2011). The position information of all genic regions are extracted from the RefFlat file downloaded from the UCSC Genome Browser (http://genome.ucsc.edu), and the putative functional impact of the variants are assessed by analyzing their positional overlap with the genic regions using an in-house Perl script. The translational impact of each small insertion is obtained from dbSNP.

The numbers of large insertion specific to each of YRI, CEU, CHB and JPT population since the CAHP are absolute numbers and may be affected by varying sample sizes. The source data from the 1000 Genome Project were generated from 59, 60, 30 and 30 individuals from YRI, CEU, CHB and JPT populations (Mills *et al.*, 2011). The 1KGP data was generated in two phases – the first phase consisted of only six individuals from two families from YRI and CEU, while the second phase consisted of all other individuals from all four populations. To avoid bias, insertions identified only in the second phase of the project were used for normalization. The number of large insertions specific to each population is normalized using the following formula:

$$I_n = \frac{nI_s}{I} \times 100$$

Where I_n is the normalized number of population specific insertion since the CAHP, n is the sample size, I_s is the raw number of population specific insertion and I is the total number of insertion identified in that population.

4.2.3 Identification of Deletion in NA18507

The whole genome sequencing (WGS) data of NA18507 were downloaded in sra format from the TRACE archive in National Center for Biotechnology Information (NCBI). The WGS data was generated using paired end sequencing (2x100bp) with ~300bp inserts at ~42x coverage by Illumina (Submission ID ERA015743, Accession ID ERX009609). The sequences were aligned against Hg19 and anc1, each as a reference genome with identical parameters using BWA alignment tool (Li & Durbin, 2009). The PCR duplicates from the aligned data were removed and the remaining aligned reads were sorted using Picard (http://picard.sourceforge.net). Discordant pairs from the sorted dataset were extracted and further aligned using a more accurate commercial alignment tool Novoalign developed by Novocraft (http://www.novocraft.com). Two SV detection tools, Meerkat (Yang *et al.*, 2013) and Delly (Rausch *et al.*, 2012), were then applied on the resulting discordant paired reads against both Hg19 and anc1 with identical parameters that were recommended by the respective authors of the two tools. Positions of the resulting SVs reported by Delly were cross-checked with all microsatellites loci in the human genome for positional overlap using Perl script to filter out the false positives associated with high levels of mis-alignments in microsatellite regions. The SV data

reported by the two tools were compared with one another both manually and using in-house Perl scripts.

4.2.4 Development of an algorithm to detect direct repeats – DRFinder

The current NGS technologies provide longer sequence reads than before with the shortest reads above 50 bps. This allows better use of the SR detection technique which gives accurate breakpoint information. One of the major benefits of having nucleotide-resolution of SV junctions is that the flanking sequences can be analyzed to assess putative mechanism. The major mechanisms for SVs – NHR, NAHR or transposition – are associated with repeat sequences of specific sizes in the flanking region. DRFinder is an algorithm that can detect repeats of up to a given size within a given range of sequence from the flanking of a genomic location. The algorithm is extremely flexible as users can define almost all parameters, with the default values set for an optimal result for common situations. The algorithm is hosted at www.sourceforge.com/p/drfinder and freely available to public for download and use.

The current version of DRFinder is only for Unix system and executable in command line only. The basic parameters of the algorithm are on the following page:

Syntax:
drfinder [options] input_file output_file
Input file format: gvf or bed
Options:

-a	Genome assembly for which the input loci are provided	hg19 (for human) or pt3 (for chimp) or pa2 (for orangutan)
-r	Search repeats only in flanking or flanking+given genomic region	Default is flanking only. t for including the given region
-s	Maximum flanking region to look for repeats	Default is 300. Can be any value.
-e	BLAST E-value for the similarity between the repeats	Default is 10^{-5}
-m	Minimum percent similarity between the sequences of the repeats	Default is 90
-f	Input file format.	Default is gvf. Type "bed" for bed format.
-l	Minimum length of target genomic locus, any entry in a list of input that is below the given length will be ignored.	Default is 0 which means no filtering
-R	Set to "t" to allow reverse complements between the repeat pairs	Default is "f" (only direct repeats)
-d	maximum bp difference allowed between the distances of two repeats from the breakpoints on both sides	Default is 5
-h	Brings up the help menu	

The algorithm is packaged in a zip file, which also contains a configuration file. The configuration file must be updated with correct locations for various supporting databases before running the program for the first time.

The tool is designed for detecting the direct repeats or TSDs (for TEIs) in down-stream analysis of new SVs. Since whole genome sequencing and analysis is very accessible nowadays, this tool can be used to quickly and efficiently provide the repeat sequences around the junctions. The sequence in flanking is a prominent indicator of the underlying mechanism for a SV, thus combining with other mechanism predictive tool

like BreakSeq (Lam *et al.*, 2010), DRFinder is an effective tool for predicting the mechanism of all SVs that may be detected in whole genome sequence analysis. An example of use of DRFinder is demonstrated in Section 4.2.3.

4.2.5 Making anc1 available to public online

Construction of an accurate genome sequence of the most recent CAHP is a continuous process. Anc1 is the first assembly of the genome sequence using all currently available data. This assembly can be improved with time as more sequencing data and more SVs become available. A website has been developed and hosted at http://genomics.brocku.ca/AncestralGenome that will serve as the central repository of the CAHP genome sequence project.

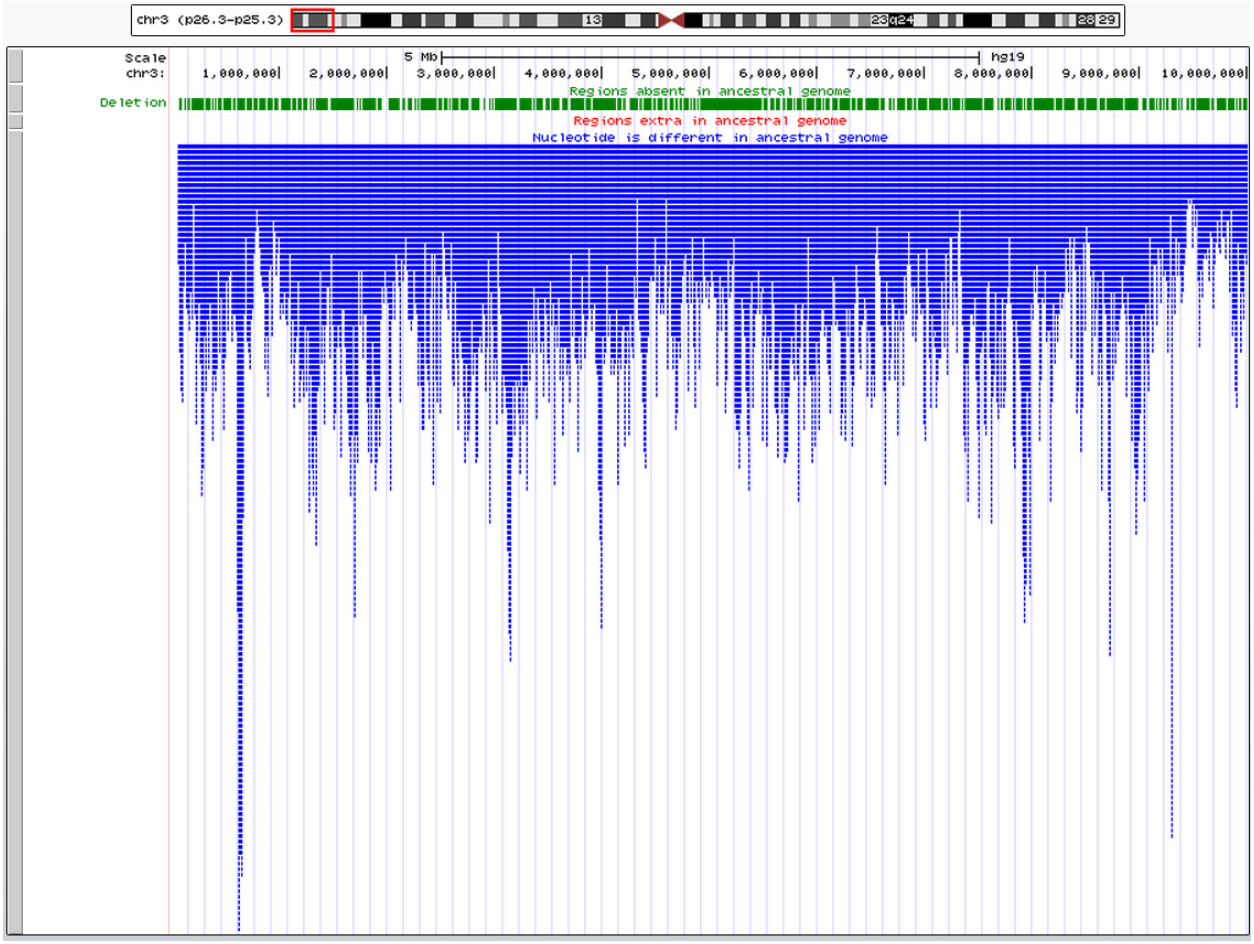
A computational pipeline has been developed to easily incorporate any new change to the genome sequence of the CAHP as more data become available. When the new changes are presented in a gvf file, the pipeline automatically compares them with the current list of changes to avoid redundancy, and then for new changes, the algorithm automatically combines them to generate the whole genome sequence of anc1. This pipeline is easy to operate and has extensive help documents so that upgrading the genome sequence of CAHP remains a continuous process and it provides a valuable resource to the human genetics and evolution research communities.

The download section of the website contains the whole genome sequence separated by each chromosome packaged in compressed files. These sequence data can be downloaded and used as the reference genome for analysis of personal genome data. The large insertions in the current reference genome that were to be deleted to obtain the

ancestral genome but could not be achieved because of lack of information for exact breakpoint are saved in a separate flat file and available for download. Furthermore, all changes from the reference genome are available for download in gvf file format. Since the most potential use of anc1 is as a reference genome instead of using Hg19 for personal genome analysis, it may be necessary for research purpose to convert Hg19 genomic coordinates into orthologous anc1 coordinates. The conversion of genomic loci between two genome assemblies can be done by Kent's tool LiftOver available from the UCSC Genome Browser. The chain files required for conversion of loci between hg19 and anc1 are also available for downloading (anc1tohg19.chain and hg19toanc1.chain). A 2bit file of the whole anc1 genome is also available for download as a database for BLAT (Kent, 2002), which is one of the fastest genome-wide alignment tools. A list of all repeat elements with their loci in anc1 is also generated using RepeatMasker and made available for download.

A genome browser hosted in UCSC is a standard tool for visualizing entire genome along with all other available associated data, such as sequence polymorphism, gene annotation, gene expression, splicing variants, and so on. The browser provides interactive visualization of the sequence and numerous controls to modify the viewing options. Almost all annotations, termed as *tracks*, for the entire genome can be visualized by turning them on or off. We have processed anc1 to make a custom track for the browser and integrated that in the website for easy representation of ancestral orthologous regions of a given genomic region in the reference genome (Figure 4.2). Users can also align a given sequence to the reference genome and see how anc1 differs in the aligned region using the browser.

a)



b)

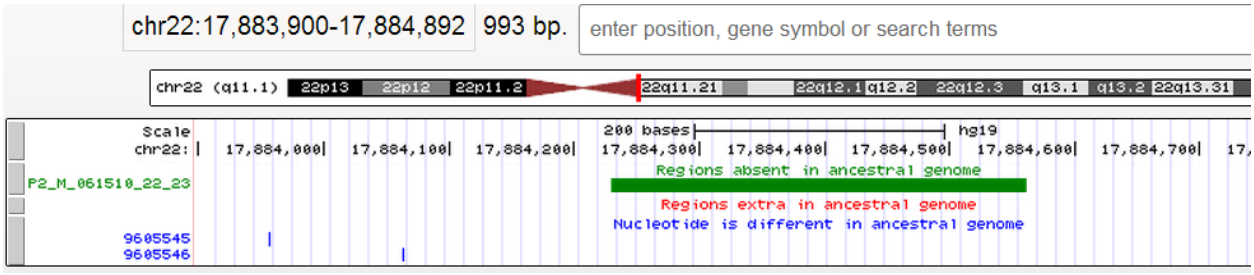


Figure 4.2 A screenshot of the genome browser for a random location in the reference genome. The green bar indicates the regions that are absent in the genome sequence of CAHP, the blue bars indicate that the nucleotides at these positions are different in the CAHP than in the reference genome. Panel a is a screenshot of a larger area containing numerous SNVs and SVs, while panel b is a screenshot of a smaller region (993 bp) that contains a deletion in anc1 and two SNVs.

4.2.6 Comparison of TE insertions between Neanderthals, the CAHP and the current reference genome

The alignment information for Neanderthal whole genome sequences against the reference genome assembly Hg19 was retrieved to our local server from the project website hosted at the Max-Planck Institute (http://www.eva.mpg.de/Neanderthal/). The genome was originally retrieved from a toe bone found in Denisova cave, then sequenced using Illumina Hi-Seq at an average of 50x coverage. Deletions present in Neanderthals compared to Hg19 were identified using Delly (Rausch *et al.*, 2012). Deletions smaller than 200 bp and larger than 10,000 bp were discarded from subsequent analysis. The remaining deletions were cross-checked for positional overlap with known satellite regions in Hg19 using a custom Perl script and any deletion overlapping a satellite region was removed. The satellite positions were obtained from the RepeatMasker (www.repeatmasker.org). The RepeatMasker information was also used to detect deletions that are transposable elements by comparing the position of each deletion with the positions of all TE insertions in Hg19. Any deletion in Neanderthal that overlaps with a TE insertion in the Hg19 is actually a TE insertion in Hg19 compared to Neanderthal, not a deletion in Neanderthal. The resulting list of TE insertions in Hg19 is then compared with the TE list in the genome sequence of the CAHP to identify the insertions that are absent in the CAHP, denoting that the insertion occurred after the CAHP, or present in CAHP, denoting the insertion occurred before the CAHP but not in Neanderthal.

4.3 Results and Discussions

4.3.1 A total of at least 8.89 Mbp DNA have been inserted into the human reference genome since the last common ancestral genome to all modern humans

Construction of the genome sequence of CAHP, the last Common Ancestor to all modern Human Populations, was made possible by the advancement of the sequencing techniques and availability of a vast amount of personal genome data and the genome sequences of other primates. The major types of variation among human individuals/populations that are included in this study are large and small sequence insertions/deletion, mobile element insertion, tandem repeat expansion, pseudogene variations and single nucleotide variations. Each of these variations was processed separately to detect the sequences that are inserted into or deleted or changed from the current reference genome since the CAHP. Insertions that are identified based on the ancestral genome lead to increase in size of the current reference genome and are contributed mostly by structural variation, i.e. segmental duplication, TEIs or other insertions, while the total number of deletions is very small and did not contribute significantly to the change in size of the current reference genome since the CAHP. (Table 4.2).

To identify insertions in the reference genome, we obtained a total of 36,777 loci in the reference genome from the structural variation data of the 1000 Genome Project (1KGP) that are absent in two or more individuals (termed as test genomes hereafter) (Mills *et al.*, 2011). These polymorphic regions could be a result of insertions in the reference genome or deletions in test genomes since the ancestral genome, and the only

method possible at the moment to assess their ancestral state is to compare the region with an outgroup primate genome for the presence or absence of the region in the outgroup genomes. However, among the 36,777 polymorphic regions, only 5,006 regions were reported with accurate breakpoint information, 24,118 regions were reported with a range of genomic positions as putative breakpoints, and 7,653 deletions in test genomes were reported with just the outer boundaries for possible breakpoints as determined by Read Depth techniques (Appendix II Figure 1a). Out of these 5,006 deletions with exact breakpoints, flanking sequences of 4,805 regions could be mapped to the chimpanzee reference genome, among which, 800 and 4,005 regions were found to be absent and present, respectively, in the chimp genome in the orthologous position. These 800 regions are likely to have been inserted in the reference genome after CAHP and the other 4,005 regions are possibly results of "deletion" events in the test genomes since the CAHP. For the 24,118 regions with a range of genomic positions as putative breakpoints, 17,939 regions could be mapped to the chimp reference genome, out of which, 1,856 regions were found absent in the chimp genome. The 7,653 deletions reported by Read Depth technique are processed separately in our study and 304 such deletions were also absent in the chimp. All these deletions in test genomes are actually insertions in the reference genome since the divergence of modern human populations and thus they are removed from the reference genome in order to construct the CAHP genome sequence. However, for the accuracy of the resulting ancestral genome, we decided to include only variations for which the breakpoints could be pinpointed to the base pair level. Thus for the 2,160 insertions in the reference genome, for which the exact breakpoint information is not available in the 1KGP data, we used the flanking sequences of the chimp orthologous

deletions to map back to the human reference genome and were able to identify the exact breakpoints of 725 such insertions. All insertions obtained from 1KGP data are more than 30bp in size and are categorized as large insertions in our study. We then further extended our detection to dbSNP using similar strategy to identify small regions that are inserted in the reference genome since the divergence of modern human populations. Out of 2,645,542 indels reported in dbSNP that are polymorphic for presence or absence among individuals, 420,306 loci were absent in the chimp genome, indicating that these regions are inserted in the reference genome after CAHP. After combining partially overlapping entries in dbSNP, 326,579 small insertions (30bp or less) and 366 large insertions were detected and subsequently removed from the reference genome to construct the genome sequence of CAHP. Many identified large insertions are actually mobile element insertions and thus reported only as mobile element insertions in the final result to avoid redundancy. In the end, we detected 863 non-TEI large insertions constituting 6.07 Mbp and 326,340 small insertions constituting 0.93 Mbp.

A similar strategy to detect large insertions in the reference genome was also used to detect transposable elements (TEs) that were inserted in the reference genome since the CAHP. dbRIP (Wang *et al.*, 2006b) and 1KGP (Stewart *et al.*, 2011) data are used as the sources for identifying human specific polymorphic TEs. When processed separately, out of 1,441 and 1,976 TEs, respectively, in dbRIP and 1KGP data that are present in the reference genome but absent in 2 or more individuals, 990 and 1,816 were also found absent in chimp when the flanking sequences were aligned against the chimp reference genome. After removing redundant entries and combining entries with overlapping positions, a total of 2,071 TEs are found to be inserted in the reference genome since the

divergence of all modern human populations, contributing 1.73 Mbp to the size increase in the reference genome. The list of the other types of repeat elements in the human genome, tandem repeats, was obtained from the Tandem Repeat Database (Gelfand *et al.*, 2007). A total of 44,559 tandem repeats with a length of 30 bp or more and at least 98% sequence similarity between the repeat units were mapped against the chimp reference genome and 17,097 loci were found to have 1 or more repeat units specific to human genome i.e., present in human but absent in chimp. 9,387 TRs among these are found to be parts of transposable element insertions that were previously identified as large insertions to have occurred since the CAHP and hence removed from further analysis. The remaining 7,710 loci may be explained as expansion of the repeat units in the human or deletion of repeat units in the chimp genome after the human-chimp divergence. In order to detect tandem repeat expansions that occurred in humans after the divergence of the modern human populations, each of 7,710 TR loci were compared with the large insertion data that are previously identified in this study to be present in the reference genome but absent in two or more individuals and the chimp reference genome. 31 such TR loci were found to have one or more repeat units that are polymorphic for presence or absence between individuals and also absent in the chimp, and these repeat units are subsequently removed from the reference genome to construct the genome sequence of CAHP. The large insertion data were also cross-checked for position overlapping with pseudogenes obtained from the Database of Pseudogene (www.pseudogene.org) and 22 pseudogenes are identified as polymorphic and absent in the chimp genome, contributing to a total of 2,965 large insertions.

An opposite strategy was taken to identify sequences that were deleted in the reference genome since the divergence of the modern human populations. These sequences are characterized by presence of the sequence in multiple individual humans and in the chimp genome but absent in the reference genome. Such sequences are only obtainable from large-scale comparative studies such as 1KGP. Out of the 14,004 deletions in the reference genome (assembly Hg19) identified from 1KGP data, only 97 sequences that are observed in 2 or more individual genomes are also found in chimp. After manually analyzing each of these 97 loci by aligning against chimp, orangutan and rhesus, 32 loci constituting a total of 1,570 bp are identified to be likely present in the CAHP genome but were deleted from the reference genome. A similar approach involving TEI resulted in detection of 14 TE loci that are present in chimp out of 4,109 loci that were identified in individual genome sequence(s) but absent in the reference genome. After manually checking these 14 loci, the MEs in four loci with a total size of 1,110 bp were found to be deleted in the reference genome after the divergence of the modern human populations. Sequence information for these 36 non-ME and ME deletions are obtained from the chimp reference genome and inserted back to construct the genome sequence of CAHP.

Other than structural variation between the genome sequence of CAHP and the reference genome, Single Nucleotide Variations (SNVs) were also included in the study to replace SNVs in the reference genome with their ancestral nucleotide. Over 42 million SNVs were retrieved from dbSNP for human reference genome assembly GRCh37, and each locus was compared with the orthologous nucleotide in the genome sequences of chimp, orangutan and macaque. A total of 5,654,377 SNPs had one of the alleles

observed in humans matching with the sequence in the orthologous primate sequence, which can be potentially considered as the ancestral nucleotide. The resulting ancestral alleles were compared with the list of ancestral alleles reported in dbSNP. For over 73% (4,147,451) SNPs, the ancestral allele we identified matched with the ancestral allele reported by dbSNP. For almost 23% SNPs, the ancestral alleles did not match and, for the remaining ~4% SNPs, dbSNP does not report any ancestral state. dbSNP identifies ancestral state of a SNP based on comparison with only chimp and their information was last updated in 2008, whereas the source data we used in this study are up-to-date with the latest assembly of the corresponding primate genomes, which may explain the large number of extra ancestral state assignments in our results. In other words, we were able to provide the ancestral status for a total of 5,654,377 SNPs with 1,506,926 not assigned in dbSNP.

Combining all differences identified between the reference genome and the putative genome sequence of CAHP, a total of 5,654,377 bases were converted to their ancestral state, 8.89 million bases were removed from, and 2,680 bases were inserted into the reference genome to construct the genome sequence of CAHP. The whole genome sequence assembly of CAHP is termed as “anc1” where “anc” is for ancestral and “1” denotes the assembly version. Anc1 is shorter than the current version of the human reference genome, GRCh37/Hg19, by 8.89 million nucleotides. The sequence of anc1 and related data files are freely available for visualization and download at the project website (http://genomics.brocku.ca/AncestralGenome).

Table 4.2 Abundance of various genomic variations and their contribution in total size change in the reference genome.

Variation	Data source	No. of loci	Total size (Mbp)
Large Insertion	1KGP (Mills *et al.*, 2011) and dbSNP	863	6.07
TE Insertion	dbRIP (Wang *et al.*, 2006b) and 1KGP (Stewart *et al.*, 2011)	2071	1.73
Small Insertion	dbSNP	326,340	0.93
Tandem repeats	TRDB (Gelfand *et al.*, 2007)	31	0.004
Pseudogenes	Database of Pseudogene (pseudogene.org)	22	0.16
Large Deletion	1KGP	32	-0.001
TE Deletion	1KGP	4	-0.001

4.3.2 Deletions have been rare events

Compared with the genome sequence of the CAHP, deletions have occurred in the reference genome only at 36 loci, four of which are precise excision of transposable elements. The non-TE deletions are all of very small sizes, ranging from 6 to 59 bp (Table 4.3). This number of non-TE deletions is likely much underrepresented because the detection of deleted regions in the reference genome since the CAHP requires detection of insertions in two or more individual genomes. Detecting insertions in test genomes is a complicated process and often the total inserted region cannot be ascertained due to limitation by short read length and short genomic fragments used in constructing the sequencing libraries.

Table 4.3 List of regions that are deleted in the Hg19 since the CAHP. The deleted events are supported by comparing each region to other primates.

Deletion site in Hg19	Size (bp)	CAHP	Chimp	Gorilla	Orangutan	Intron of gene
chr1:81617559	10	+	+	-	-	
chr10:32936400	55	+	+	-	+	C10orf68
chr10:74190853	8	+	+	-	-	MIR1256
chr11:7156563	55	+	-	+	-	
chr12:102284253	53	+	+	-	-	DRAM1
chr12:103954168	54	+	+	+	+	
chr12:112314982	46	+	+	-	-	
chr12:119272036	32	+	+	-	-	
chr12:70675596	25	+	+	-	-	CNOT2
chr13:72395561	27	+	+	-	-	DACH1
chr13:86304455	27	+	+	+	-	
chr13:90943360	58	+	+	-	-	
chr14:53304117	56	+	+	+	+	
chr17:46850403	42	+	+	+	-	TTLL6
chr18:11572230	51	+	+	-	-	
chr2:2523881	42	+	-	+	-	
chr2:49061376	57	+	+	-	-	
chr2:54627253	33	+	+	-	-	
chr3:140256987	53	+	+	+	-	CLSTN2
chr3:175062932	55	+	+	-	-	NAALADL2
chr3:80297845	46	+	+	-	-	
chr4:32803449	39	+	+	+	-	
chr5:79591601	24	+	+	-	-	
chr6:124276738	14	+	+	-	-	NKAIN2
chr6:40106183	36	+	+	-	-	
chr7:69996119	46	+	+	-	-	AUTS2
chr7:80632655	59	+	+	-	-	
chr8:60411514	6	+	+	-	-	
chr9:108420661	53	+	+	+	+	
chr9:114736428	52	+	+	+	+	
chr9:81945361	28	+	+	-	-	

To assess the functional impact of these deletions, we examined their genomic location in gene context. Among the 32 non-TE deleted region, 10 took place in the

introns of 10 different genes (Table 4.3). While it is interesting to see their highly biased distribution in intron regions (as opposed to inter-genic regions which are much larger in size), they are less likely to have any significant functional impact on the associated genes due to facts that there are in introns and small in size. Nevertheless, functional impact via alteration of splicing and regulation may not be completely excluded.

The four transposable elements that are absent in the reference genome but present in chimp and two or more individuals can be explained by three mechanisms – firstly, the TE is deleted in the reference genome after the divergence of different populations by a mechanism that is yet to be characterized; secondly, independent insertions of TE occurred at the same loci in chimp and individuals in which the TE is present; and lastly, the TE never got fixed despite its very old age and thus remained polymorphic throughout evolution. For three out of four deletions, the TSD sequence could be obtained (Table 4.4). Both copies of the TSD sequences for these three loci are present in other primates and the individuals that have the insertions, but the reference genome has only one copy of the TSD. In a study where some chimps were found to be missing TEs at certain loci but other chimps and all humans tested contained TEs at those loci, the chimps with the missing TEs only had one copy of the TSD at those loci while others had two copies (van de Lagemaat *et al.*, 2005). This can be related to TE deletions detected in our study. This indicates the first mechanism mentioned above of TE deletion in the reference genome since the CAHP is a possibility. However, it would be difficult to experimentally prove this is the mechanism responsible for these apparent "deletions". The second mechanism involving two independent insertions at the same site was observed by Conley *et al.* in which a SVA and a AluY element got inserted at the exact

same location in two individuals from two different families (Conley *et al.*, 2005). However, it is statistically very unlikely since the TEs in these cases are the same among the multiple human individuals and non-human primates carrying them. The third mechanism describing the possibility of the TE never getting fixed cannot be tested as this depends on ancestry analysis at a much larger scale. However, any genetic sequence shared by humans and chimps is commonly believed to be fixed in both species, and since one of the four TEs is also present in gorilla, it is extremely unlikely that these TEs were not fixed since the evolution of humans. Thus, the deletion of TE seems to be the most feasible mechanism and more research should be conducted to confirm that a precise removal of inserted TEs is not an impossible phenomenon. As more high quality NGS data become available, more TE insertions can be identified in individual genomes which may potentially lead to identification of more TE deletions, subsequently increasing the size of CAHP genome.

Table 4.4 The list of TE deletions in Hg19 since the CAHP. For one TE, the TSD could not be found. TE, Transposable Element; TSD, Target Site Duplication.

Deletion site in Hg19	TE subfamily	Size	TSD sequence	Anc1	Chimp	Gorilla
chr7:141013588	AluYk4	321	TCT	+	+	-
chr2:161952333	AluSx3	289	N/A	+	+	+
chr13:67903480	AluY	303	GGTG	+	+	-
chr17:25297333	AluY	283	AGTCATTAA	+	+	-

4.3.3 Smaller insertions are more abundant

While segmental duplication or TEIs contribute the most to the size increase of the reference genome, small sequence insertions obtained from dbSNP surpasses all other categories in term of abundance in number with 326,340 loci for a total of 0.93 million

bases. The other major classes of non-TE repeat elements (tandem repeats) are found to have caused the smallest increase in size of the modern human genome – only 3,878 bases. Only 31 tandem repeats have increased in copy number of the repeat units since CAHP. Pseudogenes are the fourth largest contributor to the size increase of the current human reference genome assembly (GRCh37) with 162,216 bases from 22 instances of insertions. Even though the insertions due to segmental duplication, transposition or pseudogenization contribute the most size increase in the reference genome, small insertions are significantly more abundant than any other variation type throughout the genome (Figure 4.3a). The high number of insertion loci of less than or equal to 30 bases indicates that insertion of smaller sequences occurred much more frequently in recent human evolution than did the insertions of more than 30 bases. Among the total of 326,340 such small insertions, single base insertion occurs most frequently, and the abundance gradually decreases as the insertion size increases (Figure 4.3b). The number of non-ME insertions of more than 30 bases is much smaller, only 863. Among these, 49.6% insertions are between 31 and 100bp and only 46 insertions are of more than 10kb and up to 823kb. The same pattern of decreasing abundance for increasing insertion size is observed for insertions of 31-100 bases (Figure 4.3c). The same pattern is observed for tandem repeats as well, in which the number of TR insertions that are less than 100bp in size is larger than those that are over 100bp in size (Appendix II Figure 2). The 31 tandem repeats that occurred since the CAHP range from 4 bases to 648 bases, but 19 of them are of size of less than 40 bases while the other 12 range from 40 to 648 bases. Similar bias towards smaller insertion was also observed in the case of mobile elements as the smallest type of transposable elements Alu (~300 bases) constitutes over 81% of

all TEIs. However, this agreement between size and abundance for TEIs is more suitably explained by the activity level and age of various classes of TEs (discussed later). The large and small non-TE and TE insertions were plotted onto the human chromosome ideogram based on their genomic positions in UCSC GRCh37 (Figure 4.4). Despite the non-homogenous distribution of large insertions and transposable elements, there seem to have no obvious hot spot for such insertions at the genome level.

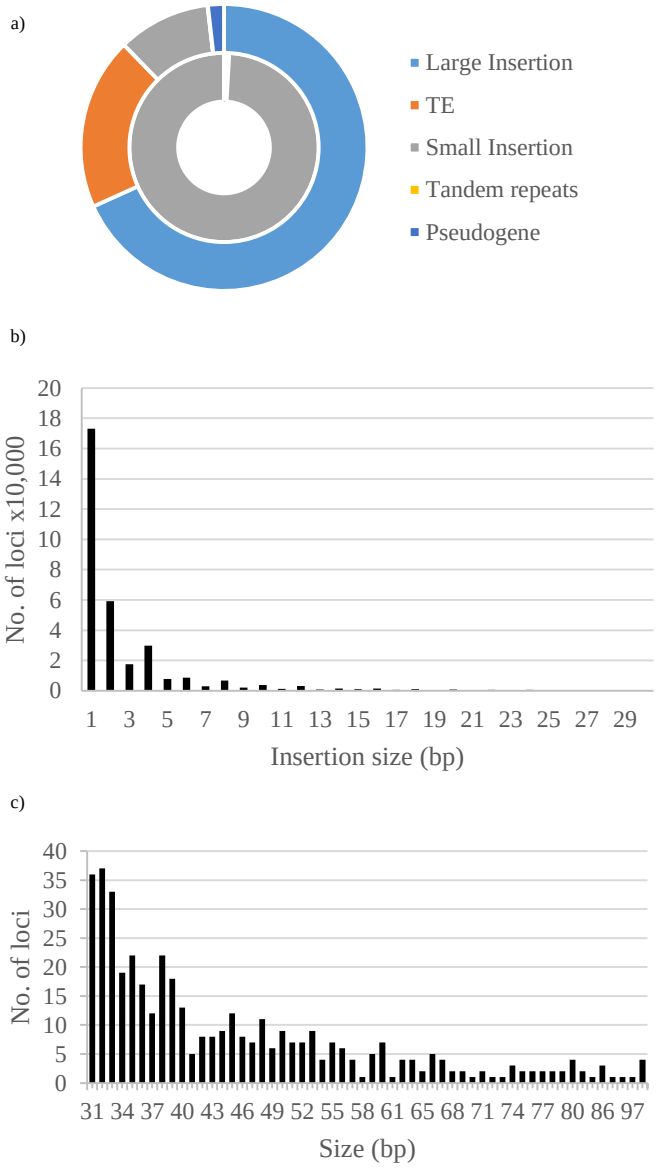


Figure 4.3 Size distribution of large and small insertions. a) Relative contribution of genomic variation towards the size increase of the current reference genome. The outer circle represents the total nucleotides contributed by each different class of variation; the inner circle represents the total number of loci. b) and c) abundance of small (1-30bp) and large (>30bp) insertions, respectively.

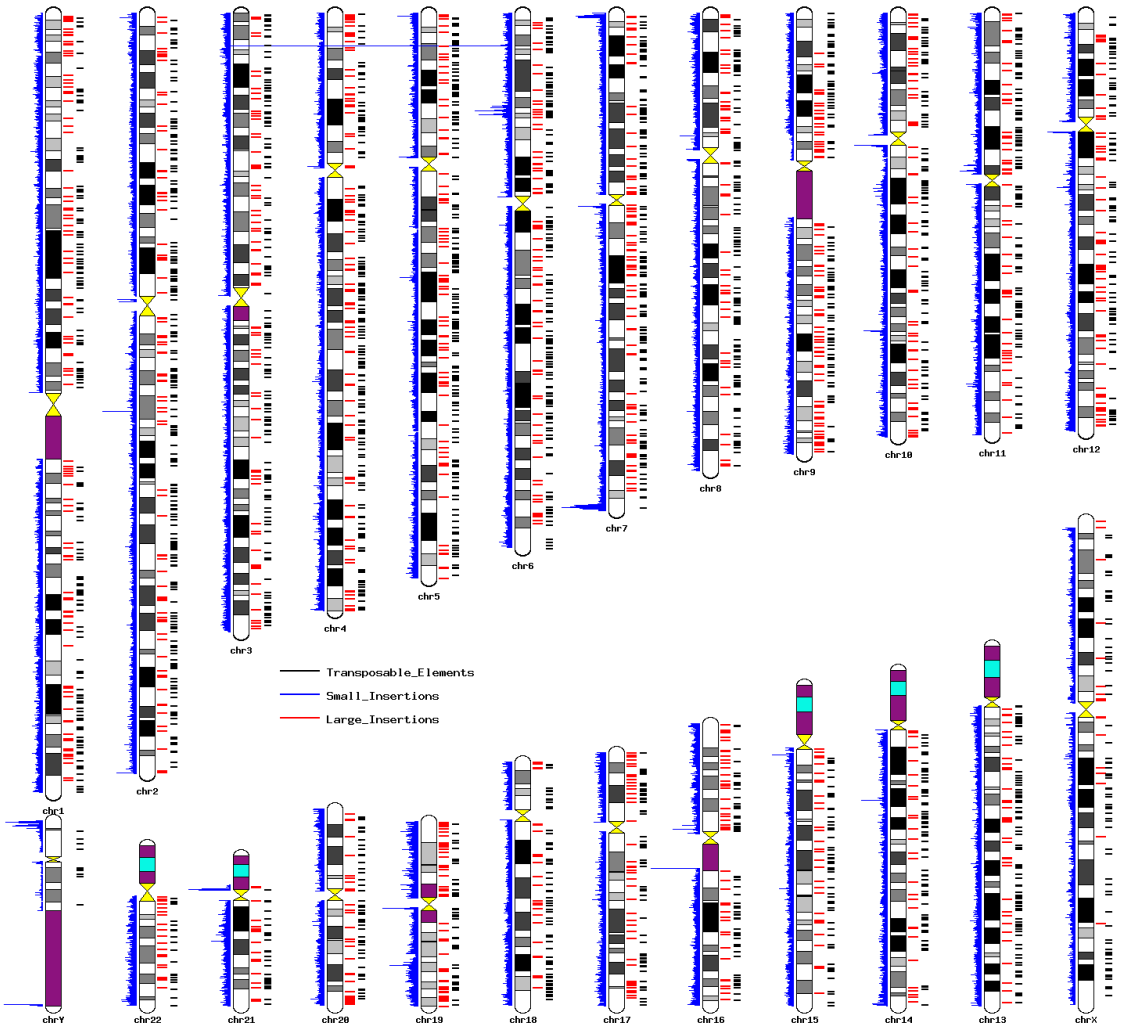


Figure 4.4 Genomic locations of all small and large insertions and transposable elements. All individual insertions (small, large and TE) are plotted onto the human chromosome ideogram based on their genomic positions in GRCh37 reference. The small insertions are denoted by blue ink on the left of the chromosomes and drawn to scale of their densities. The large insertions and transposable elements are denoted by red and black ink, respectively, on the right side of the chromosomes representing individual loci. Colored ideogram regions (mostly centromere and telomere regions) represent heterochromatin regions which mostly lack sequence information.

4.3.4 Mechanism for large sequence insertions

All large insertions were further analyzed to assess the insertion mechanism(s) because of their large contribution to the size increase of the reference genome. The assessment included all insertions over 50bp as they constitute almost 90% of the size increase as a result of segmental duplication, TE insertions (TEI), or pseudogenes that occurred in 551, 2,069, and 21 loci, respectively. The exact breakpoints resolved in this study allow us to analyze the flanking sequence of the insertions to predict the molecular mechanism. Out of 2,620 insertions (caused by duplication and TEI), 819 are flanked by direct repeats (DRs) of 6 to 296 bp of sequences within a 300 bp region on both sides of the insertions (Figure 4.5). The median length of the DRs that start within 10bp from the insertions is 12bp and the average length is 35 bp, which is a characteristic of NAHR, as well as TEI. Retrotransposition appears to be the mechanism responsible for most large sequence insertions, constituting 79% of all insertions assessed, and retrotransposon insertions are characteristically flanked by target site duplications (TSDs) of length mostly around 10bp. The high number of TEIs is largely due to abundance of Alu, L1 and SVA, which when combined constitute almost 88% of all TEI found (Figure 4.6, Appendix II Table 1). This can be explained by the fact that these major classes of TEs are still active, and as many as ~1200 Alu repeats and 147 full-length L1 insertions were predicted to be still active in previous studies (Batzer & Deininger, 2002; Mills *et al.*, 2007). The mechanism for large insertions other than TEI was also assessed using the tool breakseq (Lam *et al.*, 2010), which analyzes the junction sequence to identify homology or mechanism-specific sequence pattern in the flanking sequences. Out of the 551 non-TE large insertions that are 50 bases or longer, NAHR is found to be the dominant mechanism (~66%) (Figure 4.6). Even though transposition causes the most

number of insertion events, NAHR contributes the largest to size increase in the reference genome – 4.16 Mbp or 52% of all large insertions.

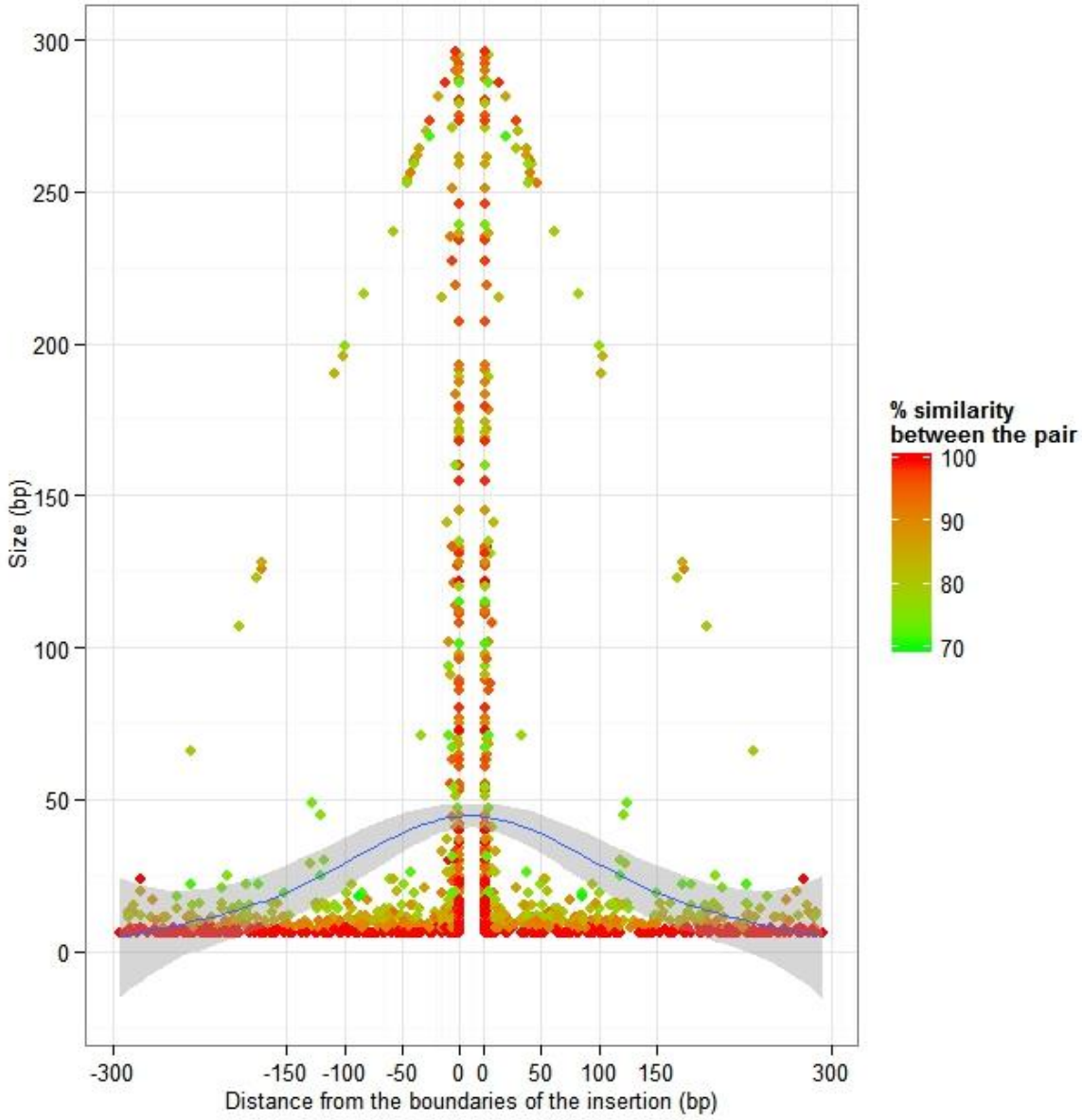


Figure 4.5 A dot plot of direct repeats surrounding 819 large insertions found in the reference genome compared to the genome sequence of the CAHP. The boundaries at each end of the insertions is marked "0" on the X-axis. The percent sequence similarity between the each pair mate is color coded with red indicating 100% similar and green indicating 70% similar. The grey curve indicates the amount of overlap of plots in the illustration, which indicates that the highest density of the repeats is closer to the boundary.

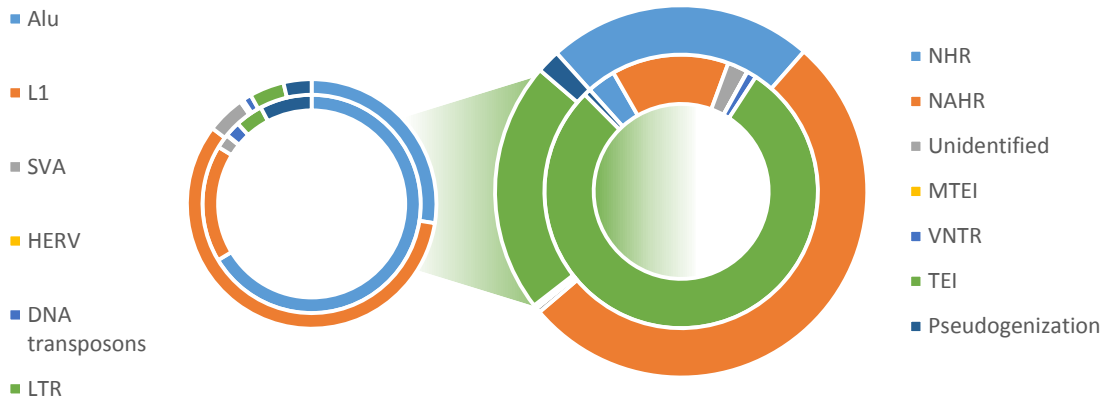


Figure 4.6 Relative abundance and contribution in size increase by different insertion mechanisms. The circles on the left side are for detailed distribution of transposable elements and those on the right are for distribution of all mechanisms involved for large insertions; the inner circle represents relative total number of loci for each different type of variation while the outer circle represents relative genomic contribution; based on a total of 2,641 insertions.

4.3.5 Analysis of genomic variants in the context of human evolution

Migration of modern humans from Africa to the rest of the world according to the "Out of Africa" theory is one of the most important feats of the recent history of modern humans. The migration gave rise to different populations at different times in history, after which, each population gained their specific genomic variations. The distribution pattern of variations across populations is very informative for understanding the migration pattern, age and genetic basis of any population-specific phenotype. In this study, we analyzed 863 large insertions resulted from segmental duplication in the reference genome sequence in 56, 57, 30 and 30 individuals from four major populations – Yoruban (YRI), Utah residents with Northern and Western European ancestry (CEU), Chinese (CHB) and Japanese (JPT) from the 1KGP project (Mills *et al.*, 2011) in an attempt to relate the insertion events with recent human evolution process. Out of the 863

insertions, population distribution for 308 insertions could be identified. It was observed that 68 (~22%) out of these 308 insertions are absent in one or more individuals from all four populations. Of the remaining 240 insertions, 47, 15, 8 and 11 insertions are found present in multiple individuals only in Yoruban, Chinese, European and Japanese populations, respectively (Figure 4.7). Since the total number of samples from each population is not the same, these numbers of population-specific large insertions are then normalized by taking sample number and total number of insertions into consideration. After normalization, it is observed that 5.17, 1.25, 0.83 and 0.73 insertions are specific to YRI, CHB, CEU and JPT populations respectively. According to the more widely accepted "Out of Africa" theory of the evolution of modern humans which states that the modern humans evolved in Africa and then migrated out to spread across the globe (Howells, 1976; Stringer & Andrews, 1988), Yoruban, Chinese, European and Japanese populations evolved chronologically. Even though the Japanese population is genetically closer to the Chinese population and they share a common early Asian ancestry, the migration to Europe took place before migration to Japan, most likely because of the sea between Chinese and Japanese lands. The insertion pattern observed in this study fits in line with this hypothesis as the older populations, Yoruban and Chinese, have more than four times more population-specific insertions than the younger ones. The absolute number of insertions specific to CEU is lower than Japanese and Chinese even though CEU evolved around 10,000 years before the Japanese population and 10,000 years after the Chinese, but this may be due to the smaller sample size for this population and corrected by normalizing the values.

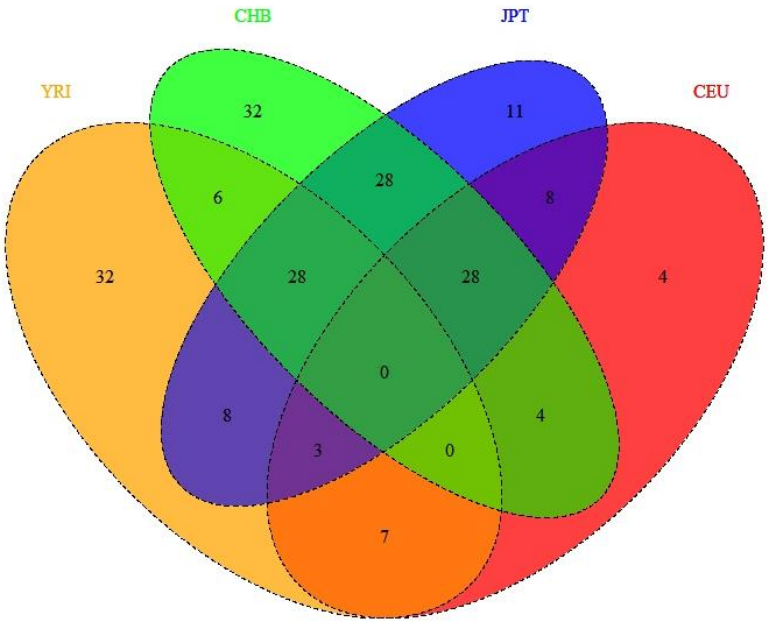


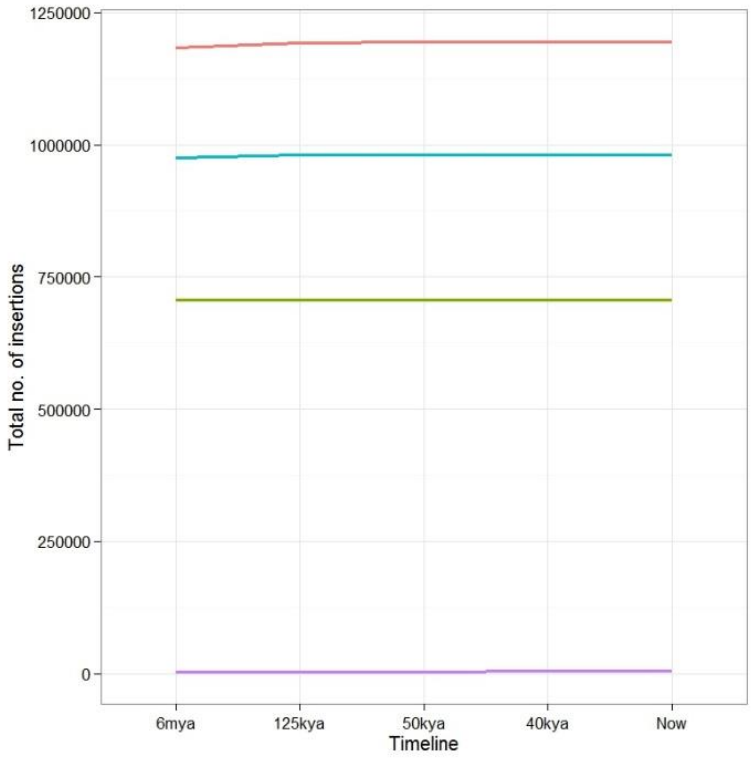
Figure 4.7 The distribution of all large insertions identified among four major populations in Venn diagram. YRI Yoruban; CHB Chinese from Beijing; JPT Japanese; CEU Residents of Utah with Central European ancestry.

Insertion of major classes of mobile elements was analyzed to identify their distribution among populations and their evolutionary expansion. For this, the allele frequencies for 1,693 TEIs among YRI, CEU and East Asian (CHB/JPT) populations were retrieved. Even though there are controversies about the exact time of evolution of different human populations, a gross assumption can be made about their relative timeframe of evolution based on their migration pattern. While the evolution of modern humans has recently been speculated to date as far back as 338 kya, most studies have indicated the evolution timeframe of modern humans before diverging into different populations to be around 125 kya ago, and subsequently migrating to China and Europe around 50 and 40 kya, respectively. A TEI with a higher allele frequency in one

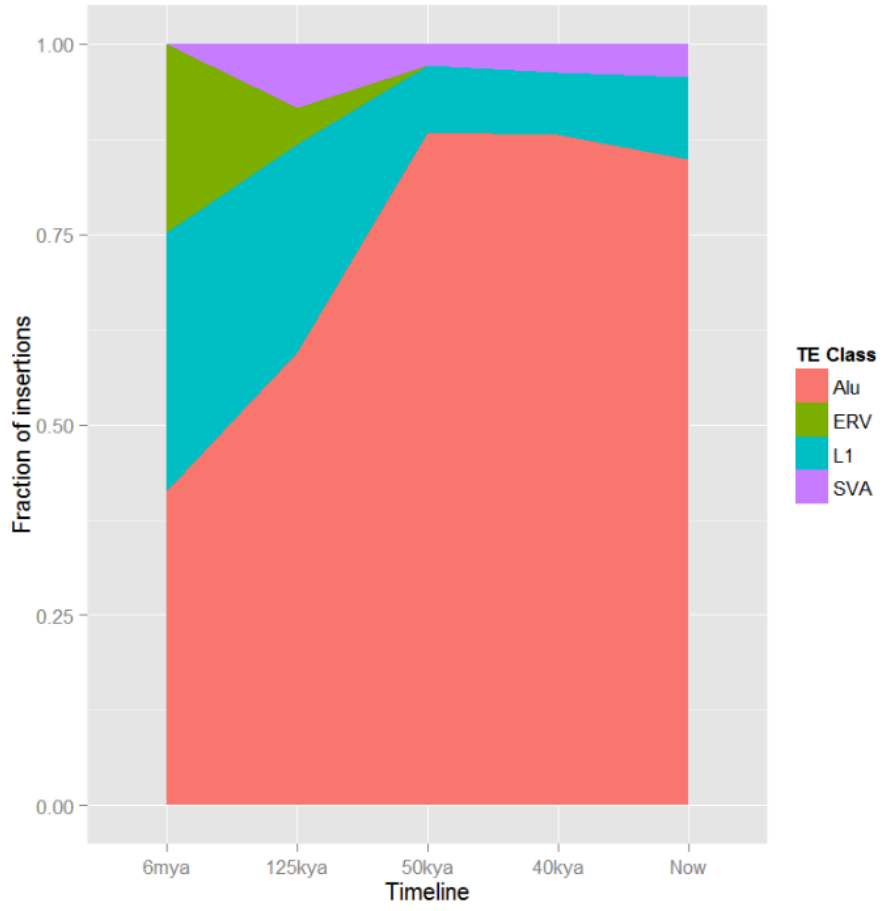
particular population can be expected to have evolved earlier during the evolution timeframe of that population, given that the chance of a growth spurt is minimal in such short time period. For example, if one insertion has a considerably higher allele frequency in Chinese population than in African or European population, we hypothesize that the earliest time of that insertion event taken place is 50 kya in aligning with the age of Chinese population. By analyzing the allele frequency of each insertion in three major human population, we estimated the rate of expansion of TEIs in human lineage at four different time period – 6 mya to 125 kya (insertions present in all human populations but absent in other primates, i.e., in CAHP), 125 kya to 50 kya (insertions dominant in African population), 50 kya to 40 kya (insertions dominant in Southeast Asian populations) and after 40 kya (insertions dominant in European populations or allele frequency too low in any population). Our analyses involving Alu, L1, SVA and ERV suggest that the growth of TEIs has been minimal and steady since the divergence of human and chimp (Figure 4.8a), with only 17,410 insertions for the period from 6 mya to CAHP around 125 kya, while the total number of insertions before 6 mya is 2,865,739. The total number of TEIs since the CAHP till today is 1,693. We further analyzed the distribution of subfamilies of MEs at each of these time periods and identified Alus to be the dominant and most active ME subfamilies compared with L1, SVA and ERV since the divergence of human and chimp (Figure 4.8b). While Alu constitutes over 80% of all insertions since human-chimp divergence, the fraction of total insertions for Alu elements has slightly gone down very recently in the evolutionary pathway. There has been minimal activity by ERV in the last 50,000 years, while the relative contribution of SVA in new insertions has gone up from almost 0% 6 mya to 3.7% now, largely because SVA

is a recently evolved family and is still actively transposing. This distribution pattern of insertions by different subfamilies of TEs is in line with their activity level and rate of expansion throughout their entire evolutionary pathway, as the total number of insertions of Alu, L1, SVA and HERV in the reference genome is in descending order and the fraction of total recent insertions occupied by these ME families are also in the same descending order. When the subclasses of Alu, L1, and SVA were analyzed, the ratio of new insertions within the subfamily is the highest for the more recent subclasses and lowest for the older subclasses in each class (Figure 4.8c) as anticipated. AluYa5 and Yb8 are the two most inserted Alu subfamilies since the CAHP. This again supports the fact that older subfamilies of TEs are lower in activity than the newer subfamilies.

a)


1250000
1000000
750000
500000
250000
0
Total no. of insertions
6mya
125kya
50kya
40kya
Now
Timeline

b)


1.00
0.75
0.50
0.25
0.00
Fraction of insertions
6mya
125kya
50kya
40kya
Now
Timeline
TE Class
Alu
ERV
L1
SVA

c)

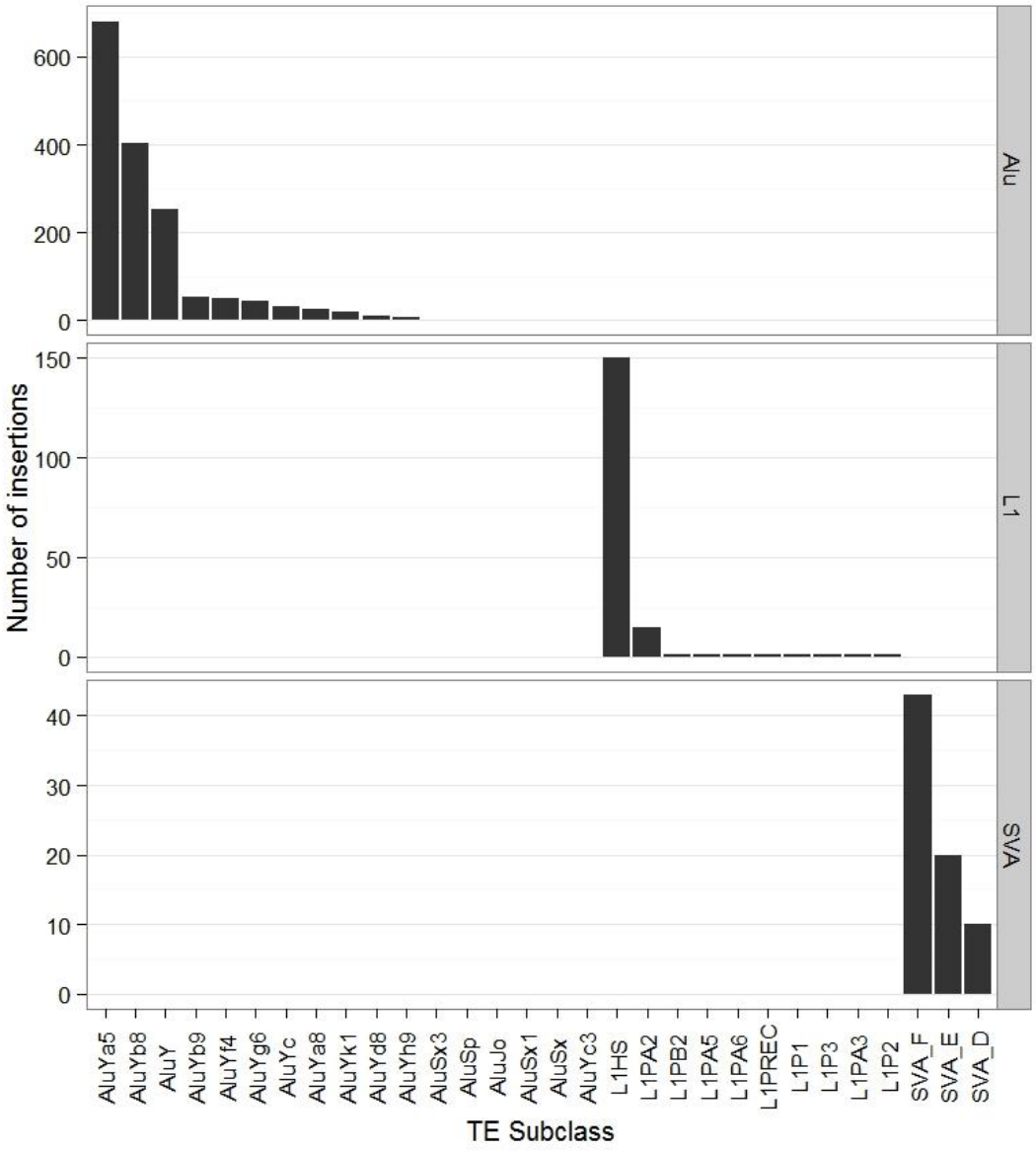


Figure 4.8 The number of transposable element insertions at different time periods in the history of modern humans. a) Numbers of new insertions of four major families of TE at different time period. b) The relative fraction of TE families that are inserted during a specific time period. c) The number of insertion of TE subfamilies in the reference genome compared to the genome sequence of CAHP.

4.3.6 Distribution of all new changes since CAHP in context of genes

Putative functional impact of all newly acquired DNA sequences in the current reference genome was assessed by comparing their genomic positions to sequence functional annotations. Almost 95% of all large insertions, 98% of small insertions, 99% of all TEs, and all TRs have occurred either in inter-genic regions or introns, and thus are less likely to have a predictable functional impact (Table 4.5). A small number of large insertions (11 cases) have caused duplication of a complete exon. All of these 11 large insertions are absent in some individuals from three or all of the four populations tested by the 1000 Genome project, indicating them as not fixed. The small number of large insertions affecting coding regions was expected, as SVs are generally less likely to overlap functional elements in the genome except for genes in certain redundant functional categories, such as receptors, ion channels, etc (Korbel *et al.*, 2007). We analyzed the function of the genes with affected coding regions using Gene Ontology (Ashburner *et al.*, 2000) and only genes from the protein domain binding category were observed to be significantly overrepresented by the affected coding sequences (with P<= 0.05 after Benjamini correction for multiple testing). After the protein domain binding category, genes with function related to receptor activity were second most overrepresented, but not to a significant extent (P=0.23). Genes of these functional categories were previously reported as more likely to harbor structural variation (Korbel *et al.*, 2007), often resulting in functional redundancy. Thus, none of these newly acquired large insertions are likely to cause any drastic functional change in the current reference genome compared to the ancestral genome.

Table 4.5 Functional impact of newly acquired sequences in the current reference genome.

Class	Gene overlap					Total number of gene	Total intergenic region
	Full coding exon overlap	Coding exon affected, partial	UTR overlap	Promoter overlap	Intron overlap		
Large insertions	11	4	11	13	290	338	525
Transposable elements	0	0	2	9	715	732	1339
Tandem repeats	0	0	0	0	13	13	18

The putative effects on coding sequences by small insertions/variations were analyzed from the data available in dbSNP (Table 4.6). Among the 120 coding sequences or reading frames that are affected by small insertions either by frame-shift, missense or nonsense mutation, only genes of protein binding functional category (GO:0005515) are significantly overrepresented (P<=0.05 after Benjamini correction for multiple testing). After the protein binding genes, the genes related with receptor binding (GO:0004872) and signal transducer activity (GO:0004871 and GO:0060089) are found to be higher in abundance among all that are affected by the small insertions, but not significantly enough to over-represent (P>0.08). Since genes involved in such molecular functions often have functional redundancy (Korbel *et al.*, 2007), changes in these coding sequences are less likely to result in any drastic change. This is very similar to the pattern observed for the large insertion. On the other hand, a significantly larger number (11,789) of coding sequences/ORFs are affected by the small sequences that are different in the reference genome than the CAHP. When these affected regions were compared to their functional categories, a significantly higher bias was observed towards protein binding, reception binding and ion binding functions (p<=0.05), which are present in many copies

in the human genome and thus loss of a few copies is unlikely to cause any drastic change. One interesting example for this is the current reference sequence for the bitterness receptor taste receptor gene, TAS2R38, carries three non-synonymous SNPs derived from the ancestral allele often as one haplotype, and individuals homozygous for this allele have no capability to taste 6-*n* propylthiouracil (PROP), a proxy for tasting bitter compounds (Bering *et al.*, 2013; Kim *et al.*, 2003). The function of this gene, related to the capability of detecting poisons in food is critical for the survival of animals, but became less important for modern humans as they became less prone to eating poison by chance. Nevertheless, this example demonstrates very well the usefulness of using the ancestral genome as the reference to more accurately assess the functional impact of genomic variants.

Table 4.6 Functional impact of small sequence variations (less than 30bp) in the current reference genome.

Type of event	Effect on coding sequence	No. of events	No. of CDS/orf	Overlapping introns	OMIM	Intergenic region
Small insertions	non-synonymous frameshift	171	134	27	6	298,110
	non-synonymous missense	1	0	1	0	
	non-synonymous nonsense	2	2	0	0	
	Synonymous	8	1	7	1	

Besides large and small insertions that overlap with some functional sequence in the genome, 21 new pseudogenes are also found to be inserted in the reference genome since the CAHP, indicating that these pseudogenes were inserted in the last 100,000 years.

Among the 21 newly acquired pseudogenes, eight are processed pseudogenes, which are believed to be generated by retrotransposition, eight are unprocessed and the mechanism of the rest five could not be determined (Table 4.7). Among all of these pseudogenes, 9 (~43%) pseodugenes are less than 80% of the size of their parents genes are thus more likely to be results of random segmental duplication.

Table 4.7 List of pseudogenes that are inserted in the reference genome (Hg19) since the CAHP. Only 12 pseudogenes are most 80% of the total size of their parent genes.

ID	Fraction	Identity	Class
PGOHUM00000246535	83%	90%	Duplicated
PGOHUM00000258377	30%	56%	Duplicated
PGOHUM00000244639	8%	97%	Ambiguous
PGOHUM00000259587	100%	97%	Processed
PGOHUM00000259588	100%	97%	Processed
PGOHUM00000245014	21%	88%	Duplicated
PGOHUM00000245223	100%	81%	Processed
PGOHUM00000245309	99%	67%	Processed
PGOHUM00000237328	62%	56%	Ambiguous
PGOHUM00000256810	80%	99%	Duplicated
PGOHUM00000258162	87%	57%	Processed
PGOHUM00000246852	55%	83%	Ambiguous
PGOHUM00000246940	100%	82%	Duplicated
PGOHUM00000250437	60%	51%	Ambiguous
PGOHUM00000234399	58%	71%	Duplicated
PGOHUM00000250421	96%	64%	Processed
PGOHUM00000241638	51%	62%	Duplicated
PGOHUM00000241517	10%	73%	Ambiguous
PGOHUM00000259931	100%	87%	Duplicated
PGOHUM00000236596	100%	70%	Processed
PGOHUM00000261414	82%	70%	Processed

Six pseudogenes (Table 4.7) contain the entire sequence of their parent genes and have relatively higher sequence identity than truncated pseudogenes. Upon further analysis of these six pseudogenes, three of them are found to be transcriptionally active,

while two out of these three transcriptionally active pseudogenes belong to the same gene family and encode a domain of the enzyme abhydrolase (ABHD17A) (Table 4.8). Both of these pseudogenes are originated from a single parent gene located in 19p13.3.

Table 4.8 List of pseudogenes that contain the entire sequence information of their parent genes. Three of these pseudogenes are transcriptionally active.

ID	Parent Protein	Parent Gene ID	Active?
PGOHUM00000259587	abhydrolase domain containing 17A	ENSG00000129968	No
PGOHUM00000259588	abhydrolase domain containing 17A	ENSG00000129968	No
PGOHUM00000245223	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 3, 9kDa	ENSG00000170906	Yes
PGOHUM00000246940	protein kinase, X-linked	ENSG00000183943	Yes
PGOHUM00000259931	SuperContig GL000222.1: 129,789-134,090	ENSG00000233094	No
PGOHUM00000236596	family with sequence similarity 27, member E2	ENSG00000204807	Yes

The number of deleted regions in the reference genome compared to the CAHP is relatively very small and is unlikely to cause any functional impact on the reference genome. Out of 36 non-TE and TE deletions, only 10 have been found to have taken place in introns, and all of these deletions are smaller in size compared to the size range of the large insertions (Table 4.3). No transposable elements that have been removed in the reference genome since the CAHP are found to be near any functionally important genomic regions.

4.3.7 Detection of deletions in NA18507 using Anc1 as a reference genome

The foundation of all downstream processes of the NGS data analysis is mapping the reads to a reference genome sequence. The standard has been to use the public human reference genome sequence, which was obtained by combining DNA sequences from 12 anonymous people from various populations (Lander *et al.*, 2001). This reference genome sequence works as the hub for comparing between two individuals, that is, to identify all differences between two individuals, each individual is compared to the reference genome, and then the results are compared to one another to see how much variations overlap or differ. This process fails to identify true insertions or deletions, as it can only call differences compared to the reference genome. These problems can be overcome by using the genome sequence of the most common recent ancestor of all modern humans (CAHP) as a reference point. Any genome sequence can be aligned against CAHP and then apply already established SV detection techniques in order to get a complete picture of true insertions and deletions. To demonstrate the usefulness of anc1 in SV analysis, the SV analysis for a genome sequence, NA18507, was compared between Hg19 and anc1 following a combinatorial approach involving PEM and SR.

Two of the most recent tools for detecting structural variations in NGS data are Meerkat (Yang *et al.*, 2013) and Delly (Rausch *et al.*, 2012). Detection of SVs in NGS data is a challenging and CPU-intensive task; thus an automatic computing pipeline to accurately call SVs in a newly sequenced genome using NGS is highly desirable. Both Meerkat and Delly were claimed to be accurate and sensitive by their respective authors, and both tools follow a similar approach for detection of SVs – firstly an initial screening of candidate SVs by PEM and secondly pinpointing the breakpoints by taking the SR

strategy. One of the major differences between the two tools is that Meerkat locally maps the unmapped reads or unmapped parts of reads against the reference genome as a part of its pipeline to include reads that have been reported as unmapped or partially mapped by the alignment tool (e.g., BWA), while Delly analyzes only read pairs reported as discordant or partially mapped by the alignment tool. Delly, on the other hand, provides the advantage of being applicable for both paired end or single read sequencing data, while Meerkat works only on paired end reads. Delly is also more robust than Meerkat because it can consider multiple sequencing libraries of varying insert sizes on the same run.

While both Delly and Meerkat produced accurate results in their benchmarking tests, the datasets used for benchmarking for the two tools were not same. In our study, we applied both tools on the Illumina sequencing data of a Yoruban individual (ID NA18507) using their respective recommended parameters in identical computational environments in order to do a comparative assessment of the two tools. When considering only deletions of 10 kbp or less, Delly and Meerkat reports 411 and 38 deletions, respectively, based on the reference genome (assembly Hg19). Out of these, deletions at only 15 loci are reported by both tools (Figure 4.9). Out of 411 deletions reported by Delly, 89 (~22%) are likely false positives since the breakpoints of these deletions overlap with microsatellite regions in the human genome. Taking likely false positives out of the equation, only ~5% of Delly outputs and ~39% of Meerkat outputs overlap with one another.

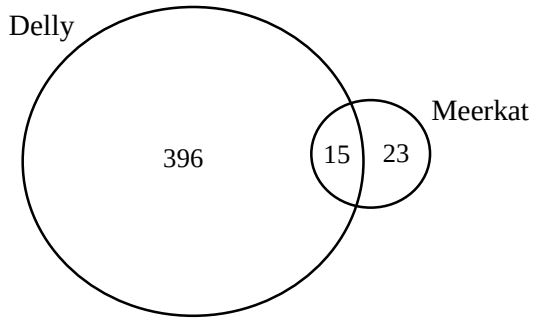


Figure 4.9 Number of deletions identified in NA18507 compared to Hg19 by two detection tools – Meerkat and Delly.

The comparative analysis of SV output by Delly and Meerkat indicates that both tools have their advantages and disadvantages in terms of accuracy and sensitivity. Both tools have high accuracy rate as presented by the benchmarking performed by their respective authors. Since further validation of deletions reports by these tools in our comparative study on NA18507 is beyond the scope of this study and since their results vary so widely, we decided to use both tools to detect deletions in NA18507 using anc1 instead of hg19 as reference sequence. Both Meerkat and Delly consistently produce fewer deletions in NA18507 when compared with anc1 than with hg19 (Table 4.9) as expected. Out of a total of 411 deletions detected using Delly with hg19 as the reference sequence, 141 (~34%) are actually insertions in hg19 and thus should be considered as insertions in hg19 rather than deletions in the test genome. Similarly, ~79% of the deletions identified by Meerkat against Hg19 are not detected against anc1, which means these deletions are actually insertions in Hg19 and not true deletions in the test genome, and thus should be considered as false positives from an evolutionary sense. Both Delly and Meerkat missed 76 and 4 deletions, respectively, when using Hg19 as the reference, because these loci are also deleted in Hg19, but got detected when using anc1 as the reference. These regions

are deleted in NA18507 but remained undetected when used Hg19 as reference genome and are thus considered as false negatives from an evolutionary context (Table 4.9). While the ancestral state of deleted regions using the conventional method could still be determined by comparison with other primates and other human individuals, using anc1 as the reference genome omits the necessity of any additional steps and produces results that are truly deletions or insertions. Furthermore, any SNVs determined using anc1 as the reference represent variations derived from the ancestral state, i.e. as new alleles.

Table 4.9 Identification of deletions of 10kb or less in NA18507 using Delly and Meerkat using both Hg19 and anc1 as reference sequences.

Tool used	Deletion reported against		False positives in conventional method	False negatives in conventional method
	Hg19	anc1		
Delly	411	346	141 (34%)	76 (22%)
Meerkat	38	12	30 (79%)	4 (30%)

Even though the difference in number of detected deletions is a good indicator of the better use of anc1 than Hg19 as a reference genome, a more useful example would be to detect the insertions in NA18507 against both genome sequences. This is because there should be significantly more insertions in NA18507 compared to anc1 than compared to Hg19, since anc1 is much smaller than Hg19 and insertions are evolutionarily more likely to take place than deletions due to higher activity by transposition and copy number gain as demonstrated by the much larger number insertions occurred from CAHP to current reference genome. However, while Delly is incapable of detecting insertions in a genome, Meerkat reported only two insertions in NA18507 compared to anc1 and no insertion compared to Hg19. A much larger number of insertions was expected compared to anc1, but the sensitivity for detection of insertions by Meerkat is low. Therefore, the

evaluation of insertion detection is not feasible with the currently available tools for this WGS dataset. This can be achieved with WGS data generated using large insert libraries or new tools with improved capability for detecting insertions.

4.3.8 TEI polymorphism between CAHP, current reference genome and Neanderthal genome

One of the major implications of the genome sequence of the most recent CAHP is in the evolutionary biology. This genome sequence represents the genetic picture of humans at a certain time period in recent human evolution, which is of particular interest because of availability of Neanderthal sequence. Even two years before, the only genome sequence we had after the divergence of human-chimp 6 mya was of the current reference genome. Now we have the genome sequence of Neanderthal, which diverged from *H. sapiens* 400,000 years ago (Hublin, 2009), and genome sequence of the most recent CAHP, which is estimated to be from 100,000 years ago, and the reference genome sequence Hg19 that represents current time. Availability of genome sequences from such varying time frame makes it feasible to study the very interesting question of how the evolution of the human genome sequence progressed, especially those genetic elements that are considered to be homoplasy-free, such as TEs.

With the alignment data of high coverage Neanderthal whole genome sequence against Hg19 (http://www.eva.mpg.de/Neanderthal/), we conducted a comparative study and identified 318 TEs that are present in Hg19 but absent in Neanderthal genome. When comparing these 318 TE insertions in Hg19 with TEs in the CAHP, the majority (189 or ~60%) of these new TEs are found to have inserted in the human lineage before the modern humans migrated out of Africa, because these TEs are also present in the CAHP

(Appendix II Table 2). The rest 129 TEs (~40%) are only present in Hg19 and absent in the CAHP and Neanderthal, thus they got inserted in the reference genome after 100,000 ya.

The distribution of TE families between those that are inserted in CAHP since Neanderthal (between 400 and 100 kya) and those that are inserted in Hg19 after the CAHP (after 100 kya) indicates that insertion of the younger and more active families constitute the majority of all families that are inserted. The most active family, Alu, has the most number of new insertions in the time between the CAHP and the current reference genome than between the Neanderthal and the CAHP. The number of insertions for L1 is much smaller than Alu and the other families have none or one insertion since CAHP. This trend is expected since Alu and L1 are two actively transposing families while L2 and other subfamilies are mostly inactive (Table 4.10). A similar trend is observed when subfamilies of Alu are analyzed (Figure 4.10). AluY, the youngest subfamily of Alu, constitutes the most insertion in the reference genome took place since the CAHP, while almost all AluS and AluJ, the older subfamilies, inserted in the modern human lineage before the CAHP.

Table 4.10 Insertion of different families of TEs in anc1 and hg19 since Neanderthal.

TE Family	No. of insertion in the CAHP since Neanderthals	No. of insertion in the Hg19 since the CAHP	No. of insertion in the Hg19 since Neanderthal
Alu	79	114	193
L1	41	14	55
ERV	21	1	22

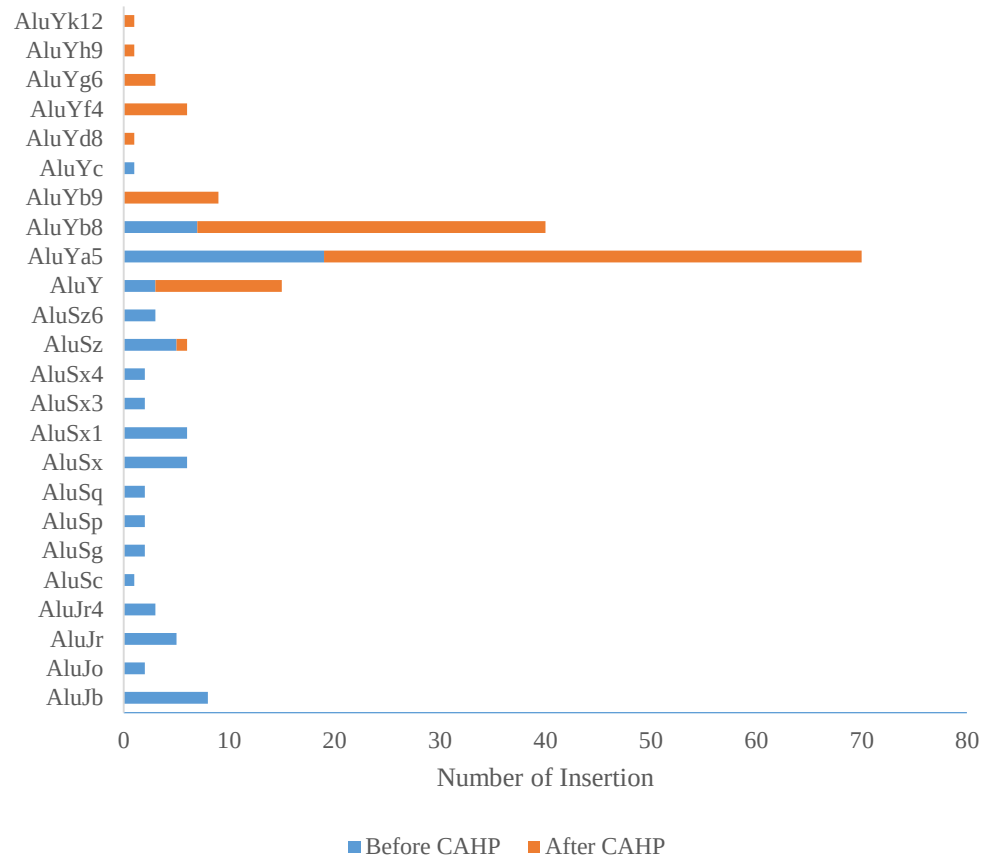


Figure 4.10 Insertion of different Alu subfamilies before and after CAHP compared with Neanderthals. The blue bar represents the fraction of total insertion of a particular Alu subfamily in the reference genome compared with Neanderthal that occurred before the CAHP, orange bar represents the insertions that occurred after the CAHP.

4.4 Conclusion

Each human individual is different from another in terms of genomic sequence composition. It is extremely important to identify all genomic variations to better understand the etiology of phenotypic variations among humans as well as to associate them with diseases or predisposition to disease. CNVs and large insertions or deletions have already been associated with multiple genetic disorders and they may potentially be

involved in causing many more disorders that are not yet well understood. The advent of sequencing technologies made large-scale whole genome a reality, which has resulted in sequencing of thousands of personal genomes to date, and many more at a magnitude of tens of thousands to come in the next few years. While the reference genome has been a standard for identification of relative differences between multiple genome samples, it fails to detect insertions and deletions in the correct evolutionary context based on the ancestral state of these genomic events. The genome sequence of CAHP proposed here is a better alternative to the reference genome that overcomes such problems. This genome sequence is also evolutionarily very important since this is the first predicted genome sequence of the common ancestor of all modern humans. This can be an extremely valuable resource in evolutionary studies representing a certain stage in the history of human evolution.

The assembly of the genome sequence of CAHP proposed here is based on the data currently available. In future, this genome assembly can be improved on multiple facets. First and foremost, it needs to include the insertions for which base-pair level sequence resolution for the breakpoints could not be resolved currently. Including Neanderthal genome sequence information during comparison with outgroup primates can help producing more accurate results. Furthermore, Neanderthal genome would provide the genomic information for another reference point during modern human evolution, and combining with CAHP and other reference genome sequences currently available, a more accurate picture of genomic evolution of modern humans can be drawn. Overall, the construction of the genome sequence of CAHP is a continuous process and the sequence will only get more accurate and reliable as more data become available.

From the initial analyses between the CAHP and current reference genome sequences, it is evident that insertions are more frequent than deletions. This phenomenon indicates that copy number gains and insertions of transposable elements are more frequently occurring events than deletions. As more polymorphic insertions, especially large tandem duplications, from the individual personal genomes get fully characterized at the detailed sequence level (they are mostly ignored for constructing anc1 due to lack of details breakpoint sequence information) by analyzing more and better personal genome data with better detection of methods, as well as identification of additional insertions based on more divergent populations, we can expect that the future version of the CAHP genome sequence of CAHP will get smaller. The detection of four transposable element removals from the reference genome is also another interesting phenomenon, as the removal of transposable elements was only observed in chimp genome before but not in humans (van de Lagemaat *et al.*, 2005). These can also be explained alternatively as old insertions that have never got fixed in human populations, and such uncertainty may be cleared by having more genome sequence data.

Chapter 5 : Overall Discussion and Conclusion

Scientific research is motivated by curiosity, but I strongly believe that it is greatly fuelled by technology and ease of operation. While the non-coding regions of the human genome including repeat elements grabbed the initial interest due to researchers' curiosity, advent of new technologies and availability of large amount of data definitely drove the research forward tremendously in the last decade. The amount of information and knowledge we have now on these genetic elements is significantly greater than any time before. More and more whole genome data are being published each year which give insights into the extent of sequence variations present among human individuals, genome-wide association among multiple elements and greater functional value to the previously undervalued genomic components, like tandem repeats or retrotransposons. My PhD thesis revolved around these genomic elements, involving extensive use of the NGS data to present novel findings related to repeat elements and a novel fundamental alternative to NGS data analyses.

NGS data enabled us to have genome sequences of individuals at very low cost. The technological advancement in recent time allowed us to have a more accurate reference genome, massive collection of personal genomics data and many databases that store gigabytes of useful genetic information. In this thesis, the availability of an accurate human reference genome and other personal genome data is utilized to analyze an important subfamily of transposable elements, AluYb. The study is them expanded to construct a relation between two types of structural variants, transposable elements and tandem repeats, for the first time in genome-wide scale utilizing the availability of several databases. The thesis also describes a systematic analysis of all variances identified to date at the individual genome sequence level, which no other single study has done

before. The approach resulted in construction of a genome sequence that represents the most recent common ancestor of all modern humans providing an extremely valuable resource to the scientific community.

Transposable elements constitute almost half of the entire human genome. Not all families of TEs are equally active with Alu elements are more active than any other families. Most of the Alu elements are thought to have been inserted in the primate genome 35-60 mya, but almost all subfamilies are inactive now. However, members of some AluY subfamilies, particularly AluYb8, are still actively retrotransposing. It is particularly interesting to study the expansion and evolutionary pattern of this young active family. TE subfamilies evolve when one or more mutation occurs in an active transposable element and that element actively transposes to generate more copies to create a new subfamily. AluYb lineage first evolved in the human lineage before humans diverged from the chimps 6 million years ago, but the majority of their insertions took place 3-4 million years ago, and thus are all human-specific. Some of these recently integrated AluYb elements retained their capability to retrotranspose. In chapter 2, the recent evolutionary history of human specific AluYb elements is analyzed using a novel approach by backtracking all full-length human-specific AluYb elements based on their sequence similarity/divergence and phylogenetic tree. One Yb8 copy is found to have generated 60% of all human-specific Yb8 elements, acting as a major driver of the expansion of Yb subfamily in the human lineage. Several other Yb8 copies are identified that generated more than 10 progenies acting as “stealth drivers” to maintain transposition capability of the subfamily in the genome. Another Yb8 copy is identified to potentially be the master copy of the Yb9 lineage. When all human specific Yb8 and

Yb9 elements were aligned with one another, multiple copies of Yb8/9 that share three different mutation patterns were detected that were previously unknown. These three sets of Yb elements that share the same mutations are classified as new subfamilies that we named as Yb8a1, Yb10 and Yb11. From an evolutionary perspective, Yb8a1, Yb10 and Yb9 are thought to have evolved within a short time span, but the total number of Yb10 identified in human reference genome is only 8 while that of Yb8a1 is 75, and that of Yb9 is almost 5 times higher than Yb8a1. This may indicate that certain mutation, e.g., Yb9-specific mutation, has increased the retrotransposition capacity in the human genome, thus the primary sequence may at least partially dictates the activity level of a TE. Furthermore, while the number of Yb10 copies is only 8 in 1.7 million years, the number of Yb11 copies is 16 in only 0.71 million years, which indicates that the Yb11 specific insertion of a T in Yb10 made the lineage more active. The association of specific mutation(s) with the activity level of TEs can be subject to future studies. The identification of new subfamilies denotes the on-going evolution of transposable elements and can be indicative of the future trend of Alu activity in the human lineage. Transposable elements function as a good marker in evolutionary studies. The homoplasy-free nature of these elements makes them critical pointers for studying evolutionary background of any individual or a population. The latest Alu subfamily, Yb11, is highly polymorphic between individuals and/or population, hence making it very useful in population genetics or studies related to migration patterns of modern humans.

In chapter 3 of this thesis, an attempt was made to construct a relationship between tandem repeats and transposable elements at the sequence level at the genome-wide scale

for the first time. The sequence and the positions of all TRs and TEs in the human reference genome were checked for similarity or overlapping among each other using computational methods. The underlying mechanisms for TRs that are suspected to have generated from TEs were also studied. It was observed that at least over 23% of all TRs in the human genome are likely to have generated from TEs constituting over 1 million base pairs. 185 of these TRs are found to be multi-locus (mlTRs) meaning that the same repeat sequence is found in more than one locus throughout the genome. For the TEs that have generated tandem repetition using a part of their sequences, younger TE families are found to be generating more TRs than the older families. Even within the same family of TEs, the younger subfamilies are observed to generate more TRs than the older subfamilies. However, the TRs generated from older subfamilies of TEs typically have higher number of repeat units than TRs generated from newer TE subfamilies. The initial unit in a tandem repeat is thought to be contributed by a TE, but the expansion of the number of repeats is mediated by *in situ* duplication rather than transposition and the number of repeats goes higher with time. We also observed that certain regions of certain TE families are more prone to generate TRs than others.

Even though the impact of transposable elements on human genome is primarily thought to be mediated by genomic rearrangement, homologous TE mediated deletion, or size increase by transposition, chapter 3 presents how they can also play an indirect role in size increase and genomic plasticity by generating TRs. The result in the chapter indicates that the number of TE-derived TRs is increasing. The study also points out the possibility of a yet-unknown mechanism for initiation and expansion of TE-derived TRs other than the mechanisms already postulated so far - slipped-strand mispairing or

replication slippage. The number of TRs derived from TEs is far more frequent than previously thought and this phenomenon should be studied more in future, especially regarding the mechanism that may lead to these events, their sequence characteristics, and functionality.

While the homoplasy-free nature of transposable elements make them a type of markers superior than SNVs for evolutionary studies, they are not the only variances present in the human genome. Chapter 4 describes a study that draws a comprehensive picture of the level of sequence variations among individual genomes using the variations data generated by Next Generation Sequencing (NGS) technologies. The recent availability of a large amount of variation data helped us to detect genome-wide structural variations, transposable element insertion polymorphism (TIP) and single nucleotide polymorphism (SNP) that are invaluable in population genetics or evolutionary studies (1000 Genomes Project Consortium *et al.*, 2010; Mills *et al.*, 2011; Stewart *et al.*, 2011). However, all of the detection methods for any of these variations are based on the publicly available human reference genome sequence. Thus, any variants detected are relative to the reference genome, and are not indicative of their true nature compared to their ancestral states. For example, an insertion detected in an individual is merely an insertion in context of the reference genome, and so it can also be understood as a deletion in the reference genome. In this case, it is not possible to obtain the evolutionary status of this variant, a piece of information necessary for assessing the functional importance of the variant. To overcome this problem and for a several other purposes, we constructed a genome sequence representative of the most recent common ancestor of all modern human populations (CAHP) by taking all genomic variations

detected among all modern human populations into consideration in their evolution history. This is the most comprehensive comparative analysis involving all major kinds of variations using all major databases as sources to propose a whole genome sequence of individual from as much as 125,000 years before. We used the latest major release of the human reference genome assembly as a base, identified all sequences that are likely inserted into or deleted from the reference genome since the evolution of CAHP, determined the ancestral state of all variants and then made necessary adjustments in the reference genome to obtain the genome sequence of the most recent CAHP. The main use of this genome sequence will be in the downstream analysis of personal genome sequencing to provide more accurate annotation of the variants based on the evolution history of the human genome. This genome sequence is also useful for identifying the gain or loss of genomic sequences and related gene function in a particular human population compared to our common ancestor. According to our results, the genome of the CAHP is at least 8.89 million bases smaller than the current reference genome. Genome sequences consensus for each modern human population can be constructed in the future and compared with the sequence of CAHP to determine how much each population differs from the common ancestor. Furthermore, this sequence is the only whole genome sequence representing humans from that time period of evolution making this a critical resource in future human evolutionary studies.

Achieving the most accurate assembly of the genome sequence of the most recent CAHP is a continuous process. The genome sequence proposed in chapter 4 is based on the data currently available. As more large scale comparative studies involving all human populations get underway, or as more or better data become available using newly

sequenced human individuals, or as better version of the reference genome becomes available, the genome sequence of the CAHP can be improved by incorporating those data. From the preliminary comparison between the genome sequences of the CAHP and the current reference genome, insertion events have been found to be predominant in recent history of modern humans. As the resolution and accuracy of the detection of variations improves, more insertions are likely to be found in the reference genome compared to the CAHP which subsequently will render the genome sequence of CAHP even smaller. Other than the improvement in sequence information, more comprehensive analyses on variable number of tandem repeats (VNTRs) or pseudogene variations may also lead to removal of more sequences from the reference genome to obtain the sequence of CAHP. Deep sequencing of individuals from the oldest population of modern humans from Africa, for example, Bantu or Khoisan, can be done to further improve the quality and validity of the genome sequence of CAHP, as the CAHP genome sequence should be more similar to these individuals than any other.

Designing and maintaining the pipeline along with all the data is extremely important to keep the improvement process of the genome sequence of the CAHP robust and continuous. Chapter 4 of this thesis also describes how the data are maintained and stored in a central online repository that is available to public. A website was developed to host all the sequence information, statistics and other necessary supporting materials for public to freely download and use in NGS data analyses. A custom track for the UCSC genome browser is also presented for easy visualization of the genome sequence of the CAHP. This genome sequence was further utilized to identify true insertions/deletions in whole genome sequence of an individual from Yoruban population as well as to identify

transposable element insertions in Neanderthals, and compared the results with insertions/deletions identified based on the reference genome. The genome data of the CAHP can also be very useful in evolutionary studies if other archaic humans can be involved, which was demonstrated in chapter 4 by involving Neanderthals to study the progression of TE expansion in recent history of modern humans.

Overall, the results presented in this thesis are valuable in the research field of transposable elements and recent evolution of modern humans. It became obvious from the studies presented here that the transposable elements have constituted a major part in the genomic changes that have occurred in very recent history of human evolution. These elements have been inserted throughout the evolution and contributed the most to the overall increase in genome size of the current modern human since the last common ancestor of all modern human populations (CAHP). The transposition events are also associated with generation of tandem repeats, which have critical role to play in genomic plasticity. While new TE insertions themselves can play functional role in the genome, possibility of the generation of new tandem repeats can create another angle by which new insertions can affect the genome. Spontaneous mutation followed by transposition can also create new subfamilies of TEs, which was observed in a study on Alu Yb subfamily presented in the thesis. The identification of three new Yb subfamilies is a critical indication of the on-going evolution of Alu elements and their future trends. The hypothesis that certain mutations in a TE subfamily can increase its activity level indicates that less-active older TE subfamilies may rejuvenate to be more active in future. Effect of mutation on activity level may also have played a part in the uneven transposition rate throughout the primate evolution and can certainly be an interesting

subject for future studies. The genome sequence of the most recent common ancestor of all modern humans that has been proposed in this thesis can not only serve as a very valuable datum for studying human evolution by providing a genome sequence of the early humans, but can also be used as a more effective alternative reference genome for the analysis of personal genome data and may change the currently established protocols of personal genome analysis processing.

Appendix I

Presented below is the supplementary information for Chapter 2: Identification of three new Alu Yb subfamilies by source tracking of recently integrated Alu Yb elements.

Table 1: List of AluYb8a1, Yb10 and Yb11 insertions identified in the reference genome.

ID	Alu Subfamily	Locus	Orientation	Alu Size
10015	AluYb10	chr5:52484718-52485036	-	318-1
10018	AluYb10	chr4:71886779-71887091	-	312-1
10019	AluYb10	chr4:86564392-86564710	-	318-1
10028	AluYb10	chr3:59386007-59386315	+	1-308
10038	AluYb10	chr20:55782867-55783184	-	317-1
10062	AluYb10	chr14:39527490-39527810	-	318-1
10067	AluYb10	chr12:12263353-12263671	-	317-1
10075	AluYb10	chr11:98476969-98477272	-	300-1
11001	AluYb11	chr9:117706509-117706828	-	318-1
11002	AluYb11	chr7:51691909-51692220	+	1-310
11003	AluYb11	chr7:133107690-133108006	+	1-315
11004	AluYb11	chr6:19155647-19155966	-	318-1
11005	AluYb11	chr6:27200713-27201032	+	1-318
11006	AluYb11	chr5:46164109-46164435	+	1-318
11007	AluYb11	chr5:119015154-119015464	-	309-1
11008	AluYb11	chr4:97572114-97572429	+	1-314
11009	AluYb11	chr2:174198592-174198905	-	312-1
11010	AluYb11	chr2:209451726-209452044	-	317-1
11011	AluYb11	chr1:70027330-70027649	+	1-318
11012	AluYb11	chr1:217754089-217754382	+	22-313
11013	AluYb11	chr14:34290048-34290372	+	1-318
11014	AluYb11	chr11:59398806-59399121	+	2-315
11015	AluYb11	chr10:104528123-104528445	-	318-1
11016	AluYb11	chr10:118664430-118664740	+	1-309
10001	AluYb8a1	chrX:38757097-38757414	+	1-317
10002	AluYb8a1	chr9:9134055-9134369	-	316-4
10003	AluYb8a1	chr9:38843977-38844285	-	308-1
10004	AluYb8a1	chr8:50304317-50304623	+	1-305
10005	AluYb8a1	chr8:144775453-144775764	-	303-1
10006	AluYb8a1	chr7:92688015-92688333	+	1-318
10007	AluYb8a1	chr7:93833428-93833745	-	318-2
10008	AluYb8a1	chr6:2788862-2789172	+	1-310
10009	AluYb8a1	chr6:67015082-67015373	-	291-1
10010	AluYb8a1	chr6:90050461-90050793	-	318-1
10011	AluYb8a1	chr6:105498195-105498512	+	1-317
10012	AluYb8a1	chr6:131863071-131863389	-	318-1
10013	AluYb8a1	chr6:137382244-137382559	+	1-318
10014	AluYb8a1	chr5:1941151-1941456	-	307-3
10016	AluYb8a1	chr5:116463743-116464055	-	309-1
10017	AluYb8a1	chr5:150315208-150315500	+	15-306
10020	AluYb8a1	chr4:113450000-113450113	-	318-206

10021	AluYb8a1	chr4:120013813-120014125	-	312-1
10022	AluYb8a1	chr4:148857468-148857779	-	310-1
10023	AluYb8a1	chr4:150554031-150554344	+	1-313
10024	AluYb8a1	chr4:175313299-175313597	-	298-1
10025	AluYb8a1	chr3:25031259-25031589	-	318-1
10026	AluYb8a1	chr3:35457965-35458283	-	318-1
10027	AluYb8a1	chr3:37327625-37327943	-	318-1
10029	AluYb8a1	chr2:20458411-20458727	-	316-1
10030	AluYb8a1	chr2:63933731-63934049	-	318-1
10031	AluYb8a1	chr2:122461322-122461637	+	1-315
10032	AluYb8a1	chr2:123830287-123830603	+	1-316
10033	AluYb8a1	chr2:212074614-212074918	-	316-1
10034	AluYb8a1	chr2:223797367-223797679	+	1-312
10035	AluYb8a1	chr21:22421331-22421597	-	304-40
10036	AluYb8a1	chr20:33338891-33339202	+	1-311
10037	AluYb8a1	chr20:35264538-35264858	+	1-316
10039	AluYb8a1	chr1:8448377-8448678	-	301-1
10040	AluYb8a1	chr1:35749561-35749876	-	315-1
10041	AluYb8a1	chr1:40635006-40635317	+	1-311
10042	AluYb8a1	chr1:161657100-161657281	+	133-313
10043	AluYb8a1	chr1:199878986-199879299	+	1-313
10044	AluYb8a1	chr1:216125187-216125505	-	318-1
10045	AluYb8a1	chr19:10441743-10442062	+	1-318
10046	AluYb8a1	chr19:46796820-46797137	+	1-317
10047	AluYb8a1	chr18:32311161-32311463	+	1-302
10048	AluYb8a1	chr18:48335225-48335543	+	1-317
10049	AluYb8a1	chr17:15202618-15202931	+	1-312
10050	AluYb8a1	chr17:36331444-36331756	-	312-1
10051	AluYb8a1	chr17:45427121-45427439	-	318-1
10052	AluYb8a1	chr16:4724622-4724940	-	318-1
10053	AluYb8a1	chr16:23638526-23638838	+	1-312
10054	AluYb8a1	chr16:72357456-72357766	+	1-310
10055	AluYb8a1	chr16:74289431-74289747	+	2-317
10056	AluYb8a1	chr15:27213440-27213756	-	316-1
10057	AluYb8a1	chr15:58282121-58282219	+	213-310
10058	AluYb8a1	chr15:66135911-66136200	+	1-289
10059	AluYb8a1	chr15:98414890-98415205	-	315-1
10060	AluYb8a1	chr14:27319580-27319897	+	1-317
10061	AluYb8a1	chr14:30415128-30415414	-	308-23
10063	AluYb8a1	chr14:40924628-40924931	+	1-303
10064	AluYb8a1	chr14:74246278-74246579	-	294-1
10065	AluYb8a1	chr14:78691093-78691408	-	315-1
10066	AluYb8a1	chr13:72506728-72507036	+	1-304
10068	AluYb8a1	chr12:66837984-66838294	+	1-310
10069	AluYb8a1	chr12:75585311-75585628	-	318-1

10070	AluYb8a1	chr12:99975386-99975705	+	1-318
10071	AluYb8a1	chr12:100832520-100832838	+	1-318
10072	AluYb8a1	chr12:120092837-120093150	-	318-1
10073	AluYb8a1	chr11:22954991-22955309	-	318-1
10074	AluYb8a1	chr11:68858148-68858434	+	30-315
10076	AluYb8a1	chr11:105765517-105765824	-	307-1
10077	AluYb8a1	chr11:123630460-123630758	+	16-313
10078	AluYb8a1	chr10:9998506-9998821	-	315-1
10079	AluYb8a1	chr10:37613923-37614159	-	307-72
10080	AluYb8a1	chr10:59053472-59053790	-	318-1
10081	AluYb8a1	chr10:88685478-88685748	-	305-36
10082	AluYb8a1	chr10:96801256-96801572	+	1-316
10083	AluYb8a1	chr10:116327667-116327977	+	1-310

Table 2: Validation of Yb11 identified outside the reference genome.

ID	Position	TE	Validated?
P1_MEI_112&P2_MEI_440	chr2:15163751-15163752	Yb11	RE in flanking
P1_MEI_3528&P2_MEI_466	chr2:218265080-218265081	Yb11	Ahmed et al.
P1_MEI_275&P2_MEI_504	chr3:111226358-111226359	Yb11	TraceDB:ti:1735507953
P1_MEI_419&P2_MEI_662	chr5:18570155-18570156	Yb11	TraceDB:ti:1734790397
P1_MEI_601	chr7:8534952-8534953	Yb11	RE in flanking
P1_MEI_894&P2_MEI_2038	chr11:13813015-13813016	Yb11	PCR failed
P1_MEI_1073&P2_MEI_2108	chr13:90947642-90947643	Yb11	RE in flanking
P1_MEI_2901&P2_MEI_143	chr13:104558081-104558082	Yb11	RE in flanking
Locus76518	chr4:150882541-150882904	Yb11	RE in flanking
Locus75350	chr4:132287226-132287647	Yb11	RE in flanking
Locus56065	chr3:84831916-84832343	Yb11	Ahmed et al.
Locus55925	chr3:81392847-81393270	Yb11	Ahmed et al.
Locus128385	chr6:165427197-165427633	Yb11	Ahmed et al.
LocusHuRef0001	chr18:53634942-53634943	Yb11	TraceDB:ti:1737279573
Locus108507	chr7:70668779-70669229	Yb11	Ahmed et al.

Table 3: List of Yb8a1, Yb10 and Yb11 insertions identified outside the reference genome. The algorithm column indicates by which algorithm the original Alu insertion was identified. RP: Paired-end method; SR: Split-read method; 1KGP: 1000 Genome Project; TE: Transposable Element; lowcov: Pilot 1 of 1KGP with low coverage genome data; trio: Pilot 2 of 1KGP with high coverage genome sequences from members of the two trio families

Position	TE	Size	Algorithm	1KGP_pilot	Reported_TE
chr1:80299098-80299099	Yb10	255	RP|SR	lowcov+trio	AluYb9
chr2:15163751-15163752	Yb11	302	RP|SR	lowcov+trio	AluYb9
chr2:218265080-218265081	Yb11	298	RP|SR	lowcov+trio	AluYb9
chr3:111226358-111226359	Yb11	304	RP|SR	lowcov+trio	AluYb9
chr4:134530754-134530755	Yb8a1	301	SR|RP	lowcov	AluYb8
chr5:18570155-18570156	Yb11	300	RP|SR	lowcov+trio	AluYb9
chr7:8534952-8534953	Yb11	301	SR|RP	lowcov	AluYb9
chr11:13813015-13813016	Yb11	304	RP|SR	lowcov+trio	AluYb9
chr12:59219044-59219045	Yb8a1	304	RP|SR	lowcov+trio	AluYb8
chr13:90947642-90947643	Yb11	299	RP|SR	lowcov+trio	AluYb8
chr13:104558081-104558082	Yb11	298	RP|SR	lowcov+trio	AluYb9
chrX:122672363-122672364	Yb8a1	302	RP|SR	lowcov+trio	AluYb8
chr4:150882541-150882904	Yb11	444	N/A	trio	AluYb9
chr4:132287226-132287647	Yb11	398	N/A	trio	AluYb9
chr4:117058410-117058822	Yb10	424	N/A	trio	AluYb9
chr1:116983311-116983773	Yb8a1	0	N/A	trio	AluYb8
chr1:77577728-77578163	Yb8a1	282	N/A	trio	AluYb8
chr14:32669080-32669331	Yb8a1	345	N/A	trio	AluYb9
chr7:64794750-64795099	Yb10	341	N/A	trio	AluYb9
chr3:84831916-84832343	Yb11	366	N/A	trio	AluYb9
chr3:81392847-81393270	Yb11	287	N/A	trio	AluYb9
chr6:165427197-165427633	Yb11	399	N/A	trio	AluYb9
chr12:105316693-105317201	Yb11	48	N/A	trio	AluYb9
chr7:70668779-70669229	Yb11	306	N/A	trio	AluYb9

Appendix II

Presented below is the supplementary information for Chapter 4: Construction of a genome sequence ancestral to all modern humans.

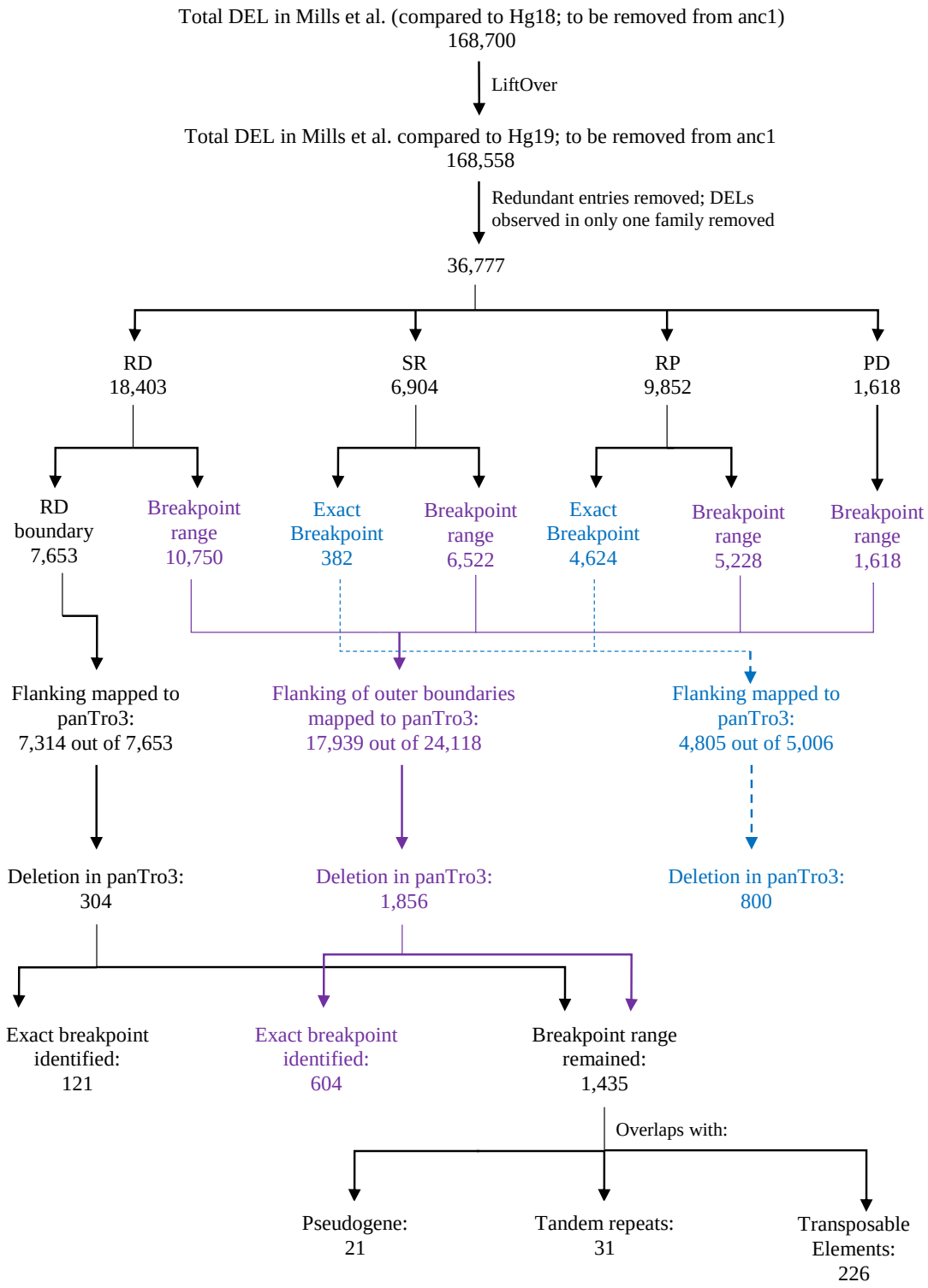


Figure 1a: A schematic flowchart of the method to detect large insertions in the reference genome compared to the genome sequence of CAHP. PD, Read depth and paired end method; RD, Read depth method; RP, Paired end method; SR, Split read method.

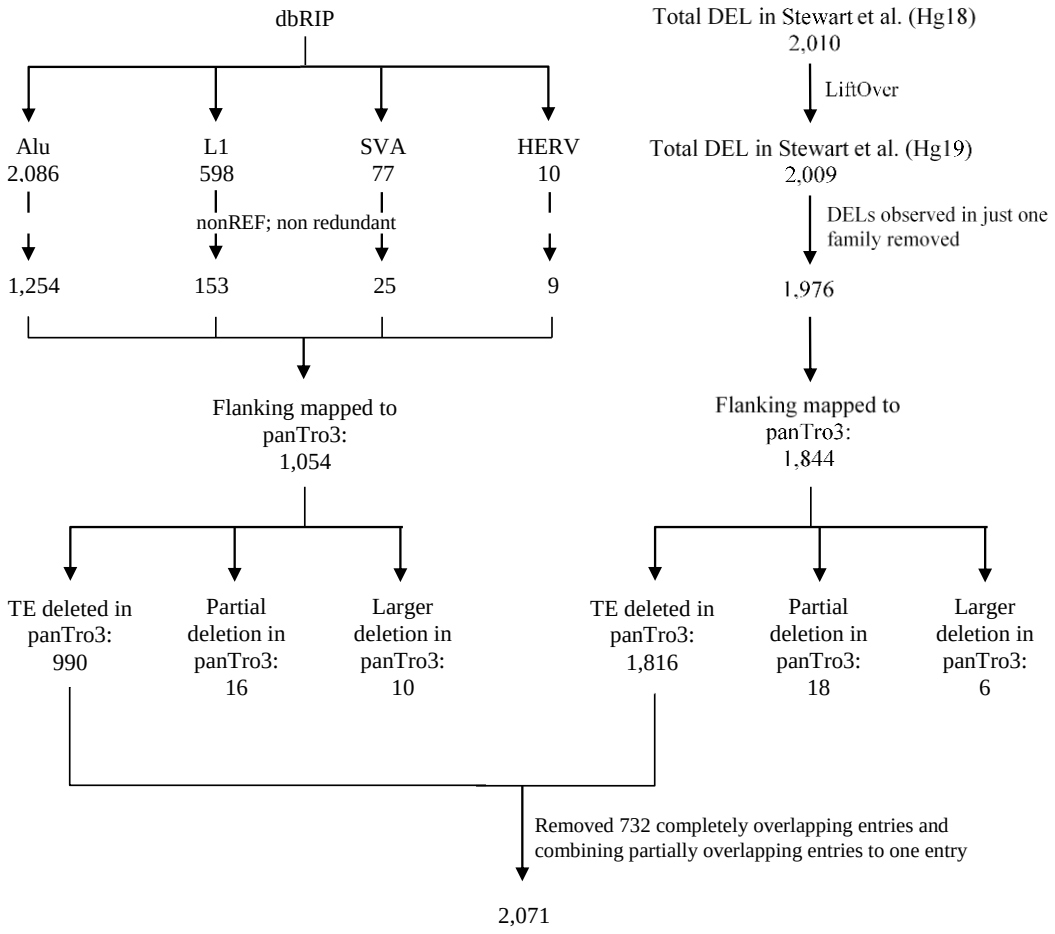


Figure 1b: A schematic flowchart of the method to transposable element insertions in the reference genome compared to the genome sequence of CAHP.

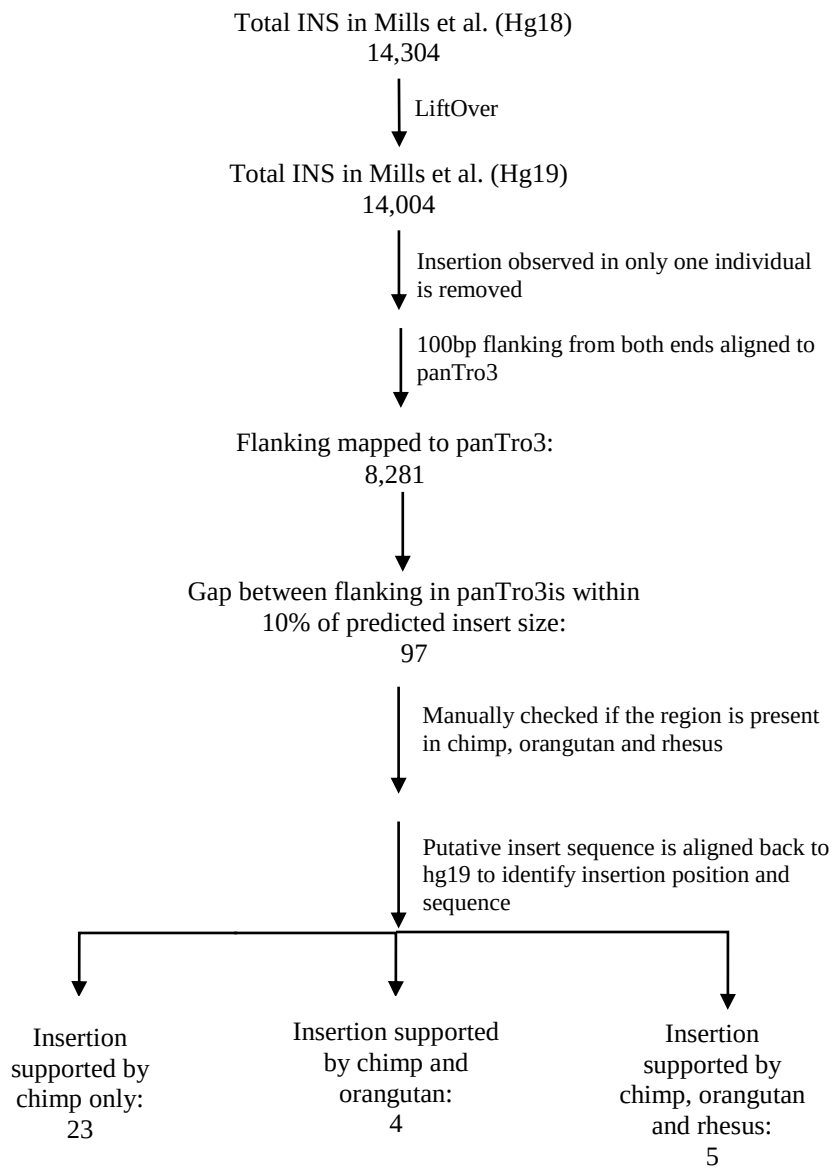


Figure 1c: A schematic flowchart of the method to detect large deletions in the reference genome compared to the genome sequence of CAHP.

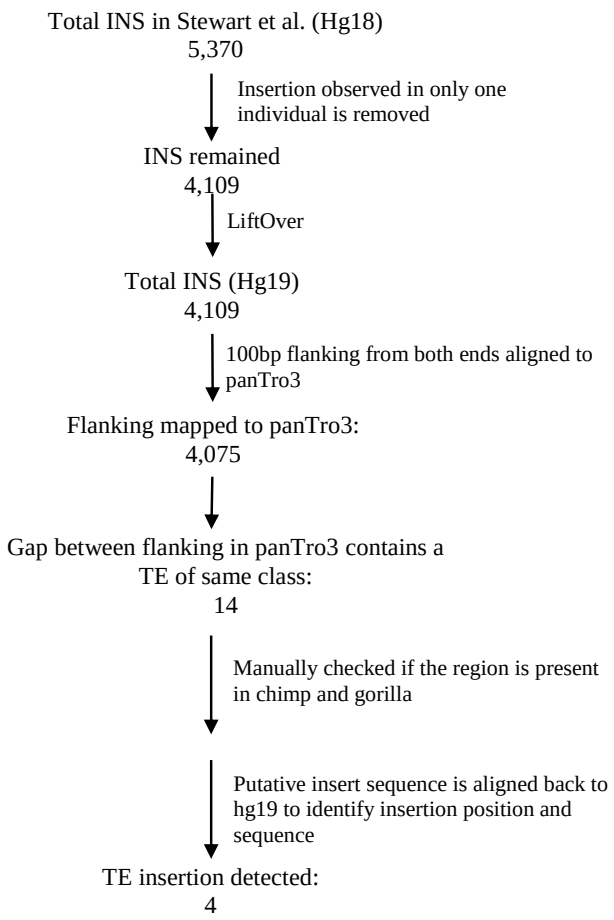


Figure 1d: A schematic flowchart of the method to detect transposable element deletions in the reference genome compared to the genome sequence of CAHP.

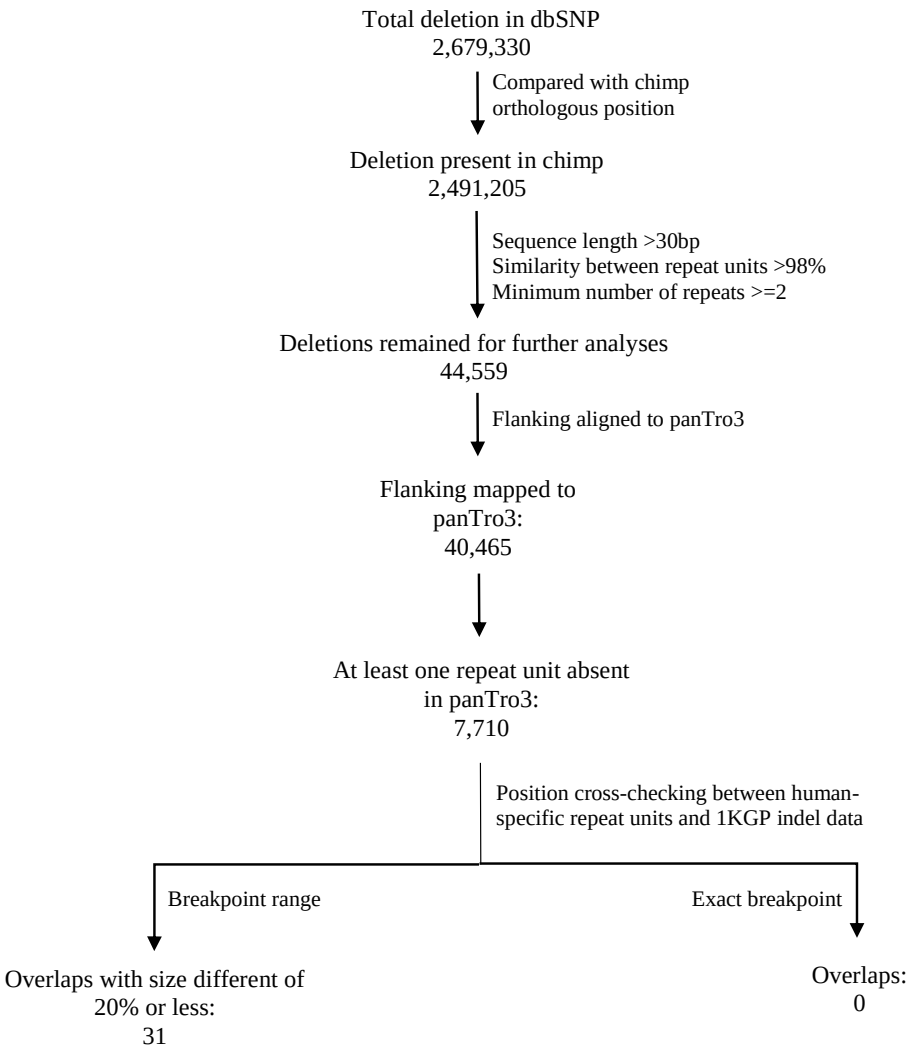


Figure 1e: A schematic flowchart of the method to detect insertions of tandem repeats in the reference genome compared to the genome sequence of CAHP.

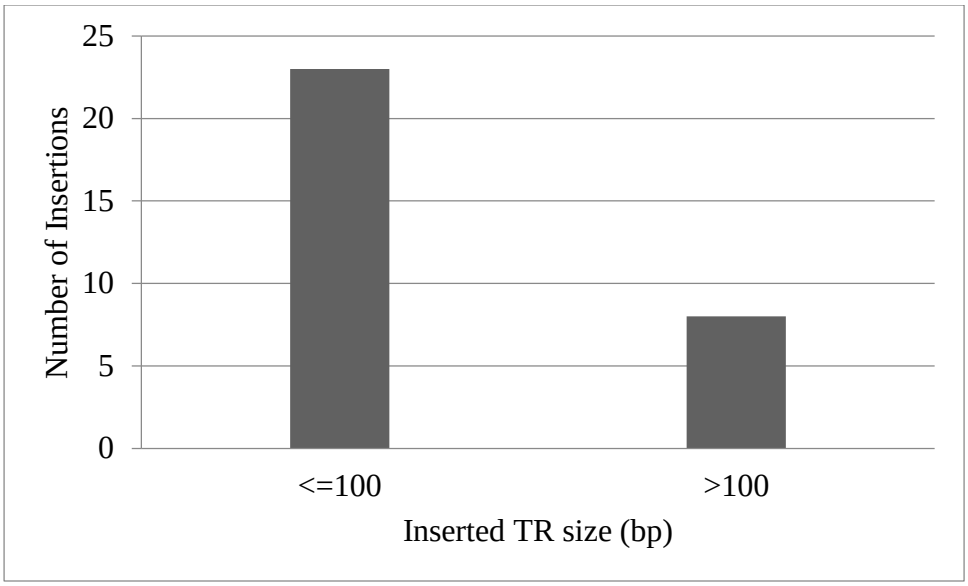


Figure 2: Number of insertions identified in GRCh37 for tandem repeats of different sizes. The TR insertions that are less than 100bp size are more abundant than TRs of more than 100 bp in size. TR, Tandem Repeats.

Table 1: Breakdown of TE classes that are inserted in the reference genome since the CAHP.

Class	Family	Subfamily	Number of Insertions
Retrotransposon	Alu	AluY	1061
Retrotransposon	L1	L1HS	127
Retrotransposon	L1	L1ME	66
Retrotransposon	Alu	AluS	60
LTR	ERV	MER	59
Retrotransposon	L1	L1PA	48
Retrotransposon	Other	SVA	41
Retrotransposon	L1	L1MA	40
LTR	ERV	MLT	40
Retrotransposon	Alu	AluJ	36
Retrotransposon	L1	L1MC	36
Retrotransposon	L1	L1MB	25
Retrotransposon	L2	L2a	19
DNA	TcMar	Tigger	18
Retrotransposon	L2	L2c	17
Retrotransposon	MIR	MIR	16
Retrotransposon	MIR	MIRb	16

Retrotransposon	L3	L3	14
Retrotransposon	L1	L1PB	13
Retrotransposon	L2	L2b	13
LTR	ERV	MER4	13
LTR	ERV	MER5	13
Retrotransposon	MIR	MIR3	13
Retrotransposon	MIR	MIRc	13
LTR	ERV	MST	13
DNA	hAT	Charlie	11
Retrotransposon	L1	HAL1	11
LTR	ERV	MER3	9
LTR	ERV	THE1	8
Retrotransposon	L1	L1M1	7
Retrotransposon	L1	L1M5	7
Retrotransposon	L1	L1MD	7
LTR	ERV	LTR8	7
Retrotransposon	ERV	HERV	6
LTR	ERV	LTR1	6
Retrotransposon	L2	L2	5
LTR	ERV	LTR4	5
LTR	ERV	MER6	5
Retrotransposon	L1	L1M3	4
Retrotransposon	L1	L1M4	4
LTR	ERV	LTR3	4
LTR	Unknown	MamRep	4
LTR	ERV	MER2	4
Retrotransposon	L1	L1P4	3
Retrotransposon	L1	L1PR	3
LTR	ERV	LTR5	3
Retrotransposon	Alu	Arth	2
Retrotransposon	ERV	ERV3	2
Retrotransposon	ERV	ERVL	2
DNA	TcMar	Kanga2	2
Retrotransposon	L1	L1M2	2
Retrotransposon	L4	L4_C_Mam	2
LTR	ERV	LTR2	2
Satellite	Centromere	ALR/Alpha	1
DNA	hAT	BLACKJACK	1
Retrotransposon	ERV	ERV2	1
Retrotransposon	Alu	FLAM	1
DNA	hAT	FordPrefect	1
Retrotransposon	L1	L1M	1
Retrotransposon	L1	L1P1	1

Retrotransposon	L1	L1P2	1
Retrotransposon	L1	L1P3	1
Retrotransposon	L1	L1P5	1
Retrotransposon	L3	L3b	1
Retrotransposon	L4	L4_A	1
rRNA		LSU	1
LTR	ERV	LTR6	1
LTR	ERV	LTR7	1
DNA	TcMar	MADE	1
LTR	Gypsy	MamGyp	1
DNA	hAT	MamTip3	1
DNA	TcMar	MARNA	1
LTR	ERV	MER7	1
LTR	ERV	MER8	1
LTR	ERV	MER9	1
LTR	ERV	PRIM	1

Table 2: Mobile element insertion polymorphism identified in Neanderthal, CAHP and Hg19.

Hg19 Position	TE Subfamily	Neanderthal	Anc1	Hg19
chr1:103047693-103048003	AluY	-	-	+
chr1:108679659-108679885	MIRb	-	+	+
chr1:113421502-113421772	L2b	-	+	+
chr1:11797560-11797732	MIRb	-	+	+
chr1:118753167-118753469	MLT1I	-	+	+
chr1:145026747-145027058	AluYb8	-	-	+
chr1:148856775-148856966	HERVL-int	-	+	+
chr1:149641220-149641436	AluSx	-	+	+
chr1:150691452-150691760	AluYa5	-	-	+
chr1:160935036-160935340	AluYa5	-	-	+
chr1:162456752-162457056	AluYa5	-	+	+
chr1:163138326-163138557	MIR	-	+	+
chr1:165730905-165731211	AluJb	-	+	+
chr1:16701811-16702107	AluSx	-	+	+
chr1:169984700-169984900	L2c	-	+	+
chr1:173944755-173945064	AluYa5	-	-	+
chr1:184315761-184315953	AluJb	-	+	+
chr1:188089924-188090114	L2a	-	+	+

chr1:22013290-22013493	HAL1	-	+	+
chr1:225134419-225134617	L1MEd	-	+	+
chr1:231303066-231303367	AluYa5	-	-	+
chr1:232423128-232423446	AluYb8	-	-	+
chr1:241149872-241150181	AluYf4	-	-	+
chr1:24402994-24403303	AluYa5	-	-	+
chr1:246923608-246923821	MIR	-	+	+
chr1:2911549-2911856	AluYg6	-	-	+
chr1:46800712-46800917	L1M5	-	+	+
chr1:57112502-57112762	L2a	-	+	+
chr1:57516583-57516771	L3	-	+	+
chr1:70027331-70027649	AluYb9	-	-	+
chr1:70768128-70768437	AluYa5	-	-	+
chr1:80218968-80219181	MIR	-	+	+
chr1:8637873-8638165	AluSx1	-	+	+
chr1:96196091-96196332	LTR33A	-	+	+
chr10:118493696-118493984	AluSx4	-	+	+
chr10:118722547-118722882	AluY	-	+	+
chr10:2116576-2116773	L1PA2	-	+	+
chr10:29709027-29709238	L1MEd	-	+	+
chr10:34153843-34154146	L2a	-	+	+
chr10:47088387-47088696	AluYa5	-	-	+
chr10:66268796-66269092	AluYa5	-	-	+
chr10:66470493-66470711	MIR	-	+	+
chr10:91547721-91548040	L1PB1	-	-	+
chr11:101416784-101417102	AluYb8	-	-	+
chr11:102603867-102604062	L1ME3G	-	+	+
chr11:125515300-125515566	L1MC4	-	+	+
chr11:19335400-19335646	AluYa5	-	+	+
chr11:20656680-20656928	MIRb	-	+	+
chr11:27307882-27308202	AluYb9	-	-	+
chr11:32286655-32286963	AluYf4	-	-	+
chr11:48354567-48354921	L1PA4	-	+	+
chr11:48446830-48447115	AluYc	-	+	+
chr11:58809303-58809619	AluYb8	-	-	+
chr11:61785949-61786150	L2b	-	+	+
chr11:62387483-62387794	AluSq	-	+	+
chr11:78416797-78416979	MIRc	-	+	+
chr11:82306896-82307205	AluYa5	-	+	+
chr11:8871400-8871679	AluJb	-	+	+
chr11:93102816-93103124	AluYf4	-	-	+
chr12:127199970-127200287	AluYb8	-	-	+
chr12:127502639-127502851	AluYb8	-	-	+

chr12:129981367-129981570	LTR33	-	+	+
chr12:26511038-26511348	AluY	-	-	+
chr12:27474591-27474863	L1MA1	-	+	+
chr12:41659832-41660137	AluYa5	-	-	+
chr12:77581362-77581561	L2c	-	+	+
chr12:77908669-77908979	AluYa5	-	-	+
chr12:82713763-82713953	L2a	-	+	+
chr13:104144135-104144379	L1MEc	-	-	+
chr13:105708104-105708423	AluYb8	-	-	+
chr13:106658153-106658333	L3b	-	+	+
chr13:111559424-111559653	AluYa5	-	-	+
chr13:20939697-20939993	AluSx1	-	+	+
chr13:23055204-23055513	AluYa5	-	-	+
chr13:44752933-44753160	L2c	-	+	+
chr13:47689491-47689791	AluYf4	-	-	+
chr13:51551947-51552263	AluYb8	-	-	+
chr13:67878731-67878952	L1PA11	-	+	+
chr13:79792046-79792245	MIRb	-	+	+
chr13:96633437-96633742	AluYh9	-	-	+
chr13:98856933-98857210	AluJr4	-	+	+
chr13:99263150-99263353	L1ME1	-	+	+
chr14:106723860-106724279	MER65A	-	+	+
chr14:27903453-27903697	L1PB4	-	+	+
chr14:73962550-73962854	AluSz6	-	+	+
chr14:76663306-76663521	MIRb	-	+	+
chr15:22415360-22415554	LTR67B	-	+	+
chr15:24369013-24369195	L1MD3	-	+	+
chr15:25388793-25389058	MER89-int	-	+	+
chr15:51897722-51898039	AluYb9	-	-	+
chr15:63269661-63269945	AluJr4	-	+	+
chr15:65776678-65776890	L1MB8	-	+	+
chr15:79891857-79892165	AluY	-	-	+
chr15:92725073-92725253	THE1B	-	+	+
chr16:16177948-16178227	AluSz	-	+	+
chr16:21452833-21453055	AluSx	-	+	+
chr16:26930631-26930834	MIRb	-	+	+
chr16:32109659-32109959	AluYa5	-	+	+
chr16:50212086-50212374	AluJr	-	+	+
chr16:62046877-62047195	AluYb9	-	-	+
chr16:63466522-63466689	L1MD1	-	+	+
chr16:70537497-70537796	AluYk12	-	-	+
chr16:74289432-74289747	AluYb8	-	-	+
chr16:75520720-75520897	L1ME3A	-	+	+

chr16:80976474-80976777	AluSx	-	+	+
chr16:84680691-84680903	AluYa5	-	+	+
chr17:13220496-13220804	AluYa5	-	-	+
chr17:34203033-34203268	MIR	-	+	+
chr17:5933615-5933924	AluYa5	-	-	+
chr17:65520867-65521184	AluYb8	-	-	+
chr17:70328469-70328721	L2c	-	+	+
chr18:27872587-27872896	AluYa5	-	-	+
chr18:34239954-34240252	AluSg	-	+	+
chr18:41453443-41453752	AluYa5	-	+	+
chr18:4207679-4207906	L1MEi	-	+	+
chr18:43261018-43261328	AluYa5	-	-	+
chr18:44338438-44338755	AluYb8	-	-	+
chr18:47870330-47876357	L1HS	-	-	+
chr18:8804244-8804412	AluJr	-	+	+
chr19:17748109-17748406	AluY	-	+	+
chr19:22768155-22768470	AluYb8	-	-	+
chr19:46741966-46742196	ERVL-E-int	-	+	+
chr19:46956108-46956413	AluYa5	-	-	+
chr19:52417325-52417582	AluJo	-	+	+
chr19:57681102-57681259	L1MA9	-	+	+
chr19:7136007-7136189	L1M4	-	+	+
chr2:113370271-113370571	AluSg	-	+	+
chr2:161795615-161795924	AluYa5	-	-	+
chr2:163666417-163666724	AluY	-	+	+
chr2:166482451-166482758	AluSq	-	+	+
chr2:168391370-168391589	MIRb	-	+	+
chr2:171793731-171793999	AluJb	-	+	+
chr2:175090641-175090950	AluYa5	-	-	+
chr2:193355413-193355707	AluYa5	-	-	+
chr2:194398509-194398818	AluYa5	-	-	+
chr2:20176387-20176582	L2c	-	+	+
chr2:206731199-206731508	AluYa5	-	-	+
chr2:209451727-209452044	AluYb9	-	-	+
chr2:213574687-213574981	AluYa5	-	-	+
chr2:214507346-214507573	AluYa5	-	+	+
chr2:227017577-227017815	AluYa5	-	-	+
chr2:233668382-233668676	AluYb8	-	-	+
chr2:4781320-4787350	L1HS	-	-	+
chr2:6167224-6167508	MLT1F1	-	+	+
chr2:64443545-64443766	MLT1O	-	+	+
chr2:72473325-72473566	L2c	-	+	+
chr2:84229904-84230131	L1MDa	-	+	+

chr2:9553454-9553758	AluSx	-	+	+
chr2:96005046-96005344	AluSz	-	+	+
chr20:18790586-18790842	MLT1L	-	+	+
chr20:26214314-26220345	L1PA2	-	+	+
chr20:29472420-29472668	L1HS	-	+	+
chr20:45211067-45211269	AluJb	-	+	+
chr21:10969157-10969445	AluYb8	-	+	+
chr21:16257078-16257395	AluYb8	-	+	+
chr21:23943933-23944127	L1MEd	-	+	+
chr21:26323491-26323718	MSTA	-	+	+
chr21:39245575-39245783	MIR	-	+	+
chr22:17677781-17678047	AluSx1	-	+	+
chr22:18047371-18047717	AluYb8	-	-	+
chr22:29177328-29177638	AluYa5	-	+	+
chr22:48740631-48740926	AluSp	-	+	+
chr3:113854032-113854347	AluYb8	-	-	+
chr3:131677834-131678143	AluYa5	-	-	+
chr3:14132685-14133652	LTR5_Hs	-	-	+
chr3:154648370-154648678	AluYa5	-	-	+
chr3:156537727-156537916	AluSx3	-	+	+
chr3:176855489-176855767	AluJb	-	+	+
chr3:188304850-188305167	AluYb8	-	-	+
chr3:191856098-191856415	AluYb8	-	-	+
chr3:25257589-25257825	L1ME3D	-	+	+
chr3:38142668-38142891	AluSx4	-	+	+
chr3:45314159-45314374	MIRc	-	+	+
chr3:51974545-51974828	AluJr	-	+	+
chr3:59520646-59520809	L1MB4	-	+	+
chr3:63851286-63851599	AluYa5	-	+	+
chr3:68870488-68870763	AluYf4	-	-	+
chr3:7347034-7347323	L1HS	-	+	+
chr3:81907508-81907817	AluYa5	-	-	+
chr3:83853136-83855348	L1HS	-	-	+
chr4:102787786-102788330	L1HS	-	-	+
chr4:103504149-103504430	AluJo	-	+	+
chr4:125671752-125672061	AluYb8	-	-	+
chr4:136454385-136454614	MIRb	-	+	+
chr4:146318834-146319151	AluYb8	-	-	+
chr4:15405931-15406128	MIRb	-	+	+
chr4:163863166-163863475	AluYa5	-	+	+
chr4:167677047-167683059	L1HS	-	-	+
chr4:168242840-168242995	ERVL-B4-int	-	+	+
chr4:171888871-171889179	AluYg6	-	-	+

chr4:178514258-178514567	AluYb8	-	-	+
chr4:182695169-182696161	L1HS	-	-	+
chr4:184754854-184755157	AluYa5	-	-	+
chr4:187795710-187796018	AluYa5	-	-	+
chr4:190253849-190254160	AluYb8	-	-	+
chr4:190624641-190624984	AluYb8	-	-	+
chr4:25497165-25497466	AluYa5	-	-	+
chr4:28971459-28971650	HAL1b	-	+	+
chr4:36469922-36470231	AluYa5	-	-	+
chr4:58148031-58148348	AluYb8	-	-	+
chr4:61035486-61035804	AluYb8	-	+	+
chr4:63487969-63488211	MIRb	-	+	+
chr4:68683103-68683320	L1M5	-	+	+
chr4:78160473-78160676	ERV3-16A3_I-int	-	+	+
chr4:81991531-81991841	AluY	-	-	+
chr4:83373236-83373429	AluJr	-	+	+
chr4:87066367-87066538	AluY	-	-	+
chr4:89523270-89523577	AluSz	-	-	+
chr4:95498973-95499280	AluYa5	-	-	+
chr4:97897221-97897531	AluYa5	-	-	+
chr5:102163372-102163608	AluYb8	-	-	+
chr5:104086992-104087168	MIR	-	+	+
chr5:104848938-104854975	L1HS	-	-	+
chr5:124607125-124607334	MIRb	-	+	+
chr5:126942345-126942659	AluYb8	-	-	+
chr5:139431920-139432154	L2a	-	+	+
chr5:149746200-149746517	AluYb8	-	-	+
chr5:151248667-151248956	AluSx1	-	+	+
chr5:165096860-165097178	AluSz6	-	+	+
chr5:172889729-172890038	AluYa5	-	-	+
chr5:20018906-20019205	AluYg6	-	-	+
chr5:20179656-20179920	LTR16A2	-	+	+
chr5:21479847-21480067	L1MA6	-	+	+
chr5:21651581-21651890	AluYa5	-	-	+
chr5:25463900-25464208	MER31A	-	+	+
chr5:33200867-33201170	AluYa5	-	-	+
chr5:3327321-3327630	AluYa5	-	+	+
chr5:36007976-36008260	AluSz	-	+	+
chr5:52356178-52356413	MIRb	-	+	+
chr5:61293641-61293958	AluYb8	-	-	+
chr5:65346062-65346372	AluYa5	-	-	+
chr5:77062317-77062619	AluYd8	-	-	+

chr5:88630519-88630734	L2	-	+	+
chr5:94109260-94109622	THE1B	-	+	+
chr6:10070806-10071105	AluYa5	-	-	+
chr6:104183196-104183505	AluYa5	-	-	+
chr6:105491020-105491236	AluSz	-	+	+
chr6:10848211-10848524	AluSx1	-	+	+
chr6:112203868-112204156	AluY	-	-	+
chr6:12617649-12617944	HAL1	-	-	+
chr6:12664202-12664510	AluYa5	-	-	+
chr6:131394531-131394770	L1M5	-	+	+
chr6:133341856-133347885	L1HS	-	-	+
chr6:147037446-147037764	AluYb8	-	-	+
chr6:159666775-159667084	AluY	-	-	+
chr6:170038131-170038447	AluYb8	-	+	+
chr6:28190215-28190532	AluYb8	-	-	+
chr6:62053990-62054299	AluYa5	-	+	+
chr6:72530095-72530293	L2c	-	+	+
chr6:85318155-85324181	L1HS	-	-	+
chr6:85574201-85574431	L1PA4	-	+	+
chr6:92808009-92808320	AluY	-	-	+
chr6:95658573-95658874	L1MA7	-	+	+
chr7:109394647-109394856	AluJb	-	+	+
chr7:110542699-110543012	AluYa5	-	-	+
chr7:112497135-112497445	AluYb8	-	+	+
chr7:113416178-113422207	L1HS	-	-	+
chr7:113953519-113953823	AluYa5	-	-	+
chr7:117548599-117548800	MIR	-	+	+
chr7:127766425-127766728	AluY	-	-	+
chr7:154671009-154671328	AluYf4	-	-	+
chr7:27514349-27514616	AluYa5	-	-	+
chr7:31946316-31946534	MIRb	-	+	+
chr7:37998512-37998822	AluYb8	-	-	+
chr7:65960038-65960206	AluJb	-	+	+
chr7:67778332-67778603	MER51B	-	+	+
chr7:67790731-67790941	AluJr4	-	+	+
chr7:7542323-7542611	AluY	-	-	+
chr7:76592351-76592533	AluSz6	-	+	+
chr7:81728755-81729064	AluYa5	-	-	+
chr7:87669442-87669746	AluSx1	-	+	+
chr7:88069052-88069247	AluYa5	-	+	+
chr7:89598880-89599188	AluYa5	-	-	+
chr7:93416960-93422991	L1HS	-	-	+
chr7:99616833-99617128	AluSx	-	+	+

chr8:10453670-10453885	L1PA8	-	+	+
chr8:107511109-107511424	AluYb8	-	-	+
chr8:112556316-112556481	L1M4	-	+	+
chr8:119447058-119447358	AluYa5	-	-	+
chr8:140019096-140019342	L1MB3	-	+	+
chr8:52061006-52061306	AluYa5	-	-	+
chr8:57027601-57027869	AluJr	-	+	+
chr8:66571143-66571369	L2	-	+	+
chr8:71306397-71306689	AluSp	-	+	+
chr8:80850477-80850666	MIRb	-	+	+
chr8:84551033-84551273	MIR	-	+	+
chr8:94284636-94284826	MIRc	-	+	+
chr8:97235553-97235733	MIRb	-	+	+
chr9:106132549-106132859	AluYa5	-	-	+
chr9:108596363-108596558	MIR	-	+	+
chr9:110432894-110433204	AluY	-	-	+
chr9:112430209-112430515	AluYa5	-	-	+
chr9:17178961-17179267	AluYa5	-	+	+
chr9:32315160-32315384	AluYa5	-	-	+
chr9:34968084-34968257	L1ME4a	-	+	+
chr9:35803106-35803400	AluYa5	-	-	+
chr9:3780267-3780574	AluY	-	-	+
chr9:38358881-38359098	LTR67B	-	+	+
chr9:72362952-72363150	L2c	-	+	+
chr9:83630453-83630756	AluYb8	-	-	+
chr9:84458135-84458444	AluYa5	-	+	+
chrX:105191163-105191463	AluYa5	-	+	+
chrX:112877030-112877339	AluYa5	-	+	+
chrX:138146446-138146697	AluYb8	-	+	+
chrX:142091770-142092036	L1PA2	-	+	+
chrX:142985537-142985955	L1M6	-	+	+
chrX:1787562-1787861	AluSx3	-	+	+
chrX:31955041-31955274	L1ME4a	-	+	+
chrX:4904719-4905022	AluYa5	-	+	+
chrX:5106606-5106868	L1M5	-	+	+
chrX:52061244-52061446	LTR16B2	-	+	+
chrX:52961292-52961607	AluYb8	-	+	+
chrX:64231583-64231877	AluSc	-	+	+
chrX:80720404-80720605	L1MA6	-	+	+
chrX:87743061-87743355	AluYb8	-	-	+
chrX:90738046-90738333	AluSz	-	+	+
chrX:91610145-91610445	AluYa5	-	+	+
chrX:97341552-97341800	MIR	-	+	+

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M. et al. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073.

Alkan, C., Coe, B. P., Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews.Genetics, 12*(5), 363-376.

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J. E., White, J. et al. (2006). An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature, 442*(7100), 295-298.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Ames, D., Murphy, N., Helentjaris, T., Sun, N., Chandler, V. (2008). Comparative analyses of human single- and multilocus tandem repeats. *Genetics, 179*(3), 1693-1704.

Arcot, S. S., Fontius, J. J., Deininger, P. L., Batzer, M. A. (1995). Identification and analysis of a 'young' polymorphic Alu element. *Biochimica Et Biophysica Acta, 1263*(1), 99-102.

Armour, J. A., Anttinen, T., May, C. A., Vega, E. E., Sajantila, A., Kidd, J. R. et al. (1996). Minisatellite diversity supports a recent African origin for modern humans. *Nature Genetics, 13*(2), 154-160.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25*(1), 25-29.

Babcock, M., Pavlicek, A., Spiteri, E., Kashork, C. D., Ioshikhes, I., Shaffer, L. G. et al. (2003). Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Research, 13*(12), 2519-2532.

Bandelt, H. J., Forster, P., Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution, 16*(1), 37-48.

Batzer, M. A., Deininger, P. L. (1991). A human-specific subfamily of Alu sequences. *Genomics, 9*(3), 481-487.

Batzer, M. A., Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews.Genetics, 3*(5), 370-379.

Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M. et al. (1996). Standardized nomenclature for Alu repeats. *Journal of Molecular Evolution, 42*(1), 3-6.

Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E. P., Stern, J. D. et al. (1995a). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *Journal of Molecular Biology, 247*(3), 418-427.

Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E. P., Stern, J. D. et al. (1995b). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *Journal of Molecular Biology, 247*(3), 418-427.

Batzer, M. A., Schmid, C. W., Deininger, P. L. (1993). Evolutionary analyses of repetitive DNA sequences. *Methods in Enzymology*, *224*, 213-232.

Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H. et al. (1994). African origin of human-specific polymorphic Alu insertions. *Proceedings of the National Academy of Sciences of the United States of America, 91*(25), 12288-12292.

Bering, A. H., Pickering, G., Liang, P. (2013). Single Nucleotide Polymorphisms Are Associated with PROP—but Not Thermal—Tasting: a Pilot Study. *Chemosensory Perception,*

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J. et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature, 463*(7283), 899-905.

Bird, C. P., Stranger, B. E., Liu, M., Thomas, D. J., Ingle, C. E., Beazley, C. et al. (2007). Fast-evolving noncoding sequences in the human genome. *Genome Biology*, *8*(6), R118.

Boeke, J. D. (1997). LINEs and Alus--the polyA connection. *Nature Genetics, 16*(1), 6-7.

Bois, P., Jeffreys, A. J. (1999). Minisatellite instability and germline mutation. *Cellular and Molecular Life Sciences : CMLS, 55*(12), 1636-1648.

Bois, P. R. (2003). Hypermutable minisatellites, a human affair? *Genomics, 81*(4), 349-355.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature, 368*(6470), 455-457.

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prufer, K. et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America, 104*(37), 14616-14621.

Carter, A. B., Salem, A. H., Hedges, D. J., Keegan, C. N., Kimball, B., Walker, J. A. et al. (2004). Genome-wide analysis of the human Alu Yb-lineage. *Human Genomics, 1*(3), 167-178.

Chan, S. W., Zhang, X., Bernatavichute, Y. V., Jacobsen, S. E. (2006). Two-step recruitment of RNA-directed DNA methylation to tandem repeats. *PLoS Biology, 4*(11), e363.

Charlesworth, B. (1994). Genetic recombination. Patterns in the genome. *Current Biology : CB, 4*(2), 182-184.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research, 31*(13), 3497-3500.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L. et al. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods, 6*(1), 99-103.

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature, 437*(7055), 69-87.

Churakov, G., Grundmann, N., Kuritzin, A., Brosius, J., Makalowski, W., Schmitz, J. (2010). A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. *BMC Evolutionary Biology, 10*, 376-2148-10-376.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A. et al. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science (New York, N.Y.), 302*(5652), 1960-1963.

Claverie-Martin, F., Gonzalez-Acosta, H., Flores, C., Anton-Gamero, M., Garcia-Nieto, V. (2003). De novo insertion of an Alu sequence in the coding region *Human Genetics, 113*(6), 480-485.

Conley, M. E., Partain, J. D., Norland, S. M., Shurtleff, S. A., Kazazian, H. H.,Jr. (2005). Two independent retrotransposon insertions at the same site within the coding region of BTK. *Human Mutation, 25*(3), 324-325.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature, 464*(7289), 704-712.

Cordaux, R., Hedges, D. J., Batzer, M. A. (2004). Retrotransposition of Alu elements: how many sources? *Trends in Genetics : TIG, 20*(10), 464-467.

Deininger, P. L., Batzer, M. A., Hutchison, C. A.,3rd, Edgell, M. H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends in Genetics : TIG, 8*(9), 307-311.

Dewannieux, M., Esnault, C., Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics, 35*(1), 41-48.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A. et al. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature, 380*(6570), 152-154.

Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., Damron-Boles, D., Husain, Z. et al. (1996). A comprehensive genetic map of the mouse genome. *Nature, 380*(6570), 149-152.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution, 17*(6), 368-376.

Ferlini, A., Galie, N., Merlini, L., Sewry, C., Branzi, A., Muntoni, F. (1998). A novel Alu-like element rearranged in the dystrophin gene causes a splicing mutation in a family with X-linked dilated cardiomyopathy. *American Journal of Human Genetics, 63*(2), 436-446.

Feuk, L., Carson, A. R., Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews.Genetics, 7*(2), 85-97.

Finlayson, C., Pacheco, F. G., Rodriguez-Vidal, J., Fa, D. A., Gutierrez Lopez, J. M., Santiago Perez, A. et al. (2006). Late survival of Neanderthals at the southernmost extreme of Europe. *Nature, 443*(7113), 850-853.

Ganguly, S., Ghosh, A., Biswas, S., Halder, R. (2003). Laboratory investigations for rheumatic disease. *Journal of the Indian Medical Association, 101*(11), 664-666.

Gelfand, Y., Rodriguez, A., Benson, G. (2007). TRDB--the Tandem Repeats Database. *Nucleic Acids Research, 35*(Database issue), D80-7.

Gentles, A. J., Kohany, O., Jurka, J. (2005). Evolutionary diversity and potential recombinogenic role of integration targets of Non-LTR retrotransposons. *Molecular Biology and Evolution, 22*(10), 1983-1991.

Goodman, M., Porter, C. A., Czelusniak, J., Page, S. L., Schneider, H., Shoshani, J. et al. (1998). Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Molecular Phylogenetics and Evolution, 9*(3), 585-598.

Grun, R., Stringer, C., McDermott, F., Nathan, R., Porat, N., Robertson, S. et al. (2005). U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *Journal of Human Evolution, 49*(3), 316-334.

Haber, J. E., Louis, E. J. (1998). Minisatellite origins in yeast and humans. *Genomics, 48*(1), 132-135.

Hagelberg, E., Gray, I. C., Jeffreys, A. J. (1991). Identification of the skeletal remains of a murder victim by DNA analysis. *Nature, 352*(6334), 427-429.

Han, K., Xing, J., Wang, H., Hedges, D. J., Garber, R. K., Cordaux, R., Batzer, M. A. (2005). Under the genomic radar: the stealth model of Alu amplification. *Genome Research, 15*(5), 655-664.

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K. D., Wray, G. A. (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics, 39*(9), 1140-1144.

Hedges, D. J., Callinan, P. A., Cordaux, R., Xing, J., Barnes, E., Batzer, M. A. (2004). Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Research, 14*(6), 1068-1075.

Hickey, G., Paten, B., Earl, D., Zerbino, D., Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics (Oxford, England), 29*(10), 1341-1342.

Hormozdiari, F., Alkan, C., Ventura, M., Hajirasouliha, I., Malig, M., Hach, F. et al. (2011). Alu repeat discovery and characterization within human genomes. *Genome Research, 21*(6), 840-849.

Howells, W. W. (1976). Explaining Modern Man: Evolutionists versus Migrationists. *Journal of Human Evolution, 5*, 477-495.

Hublin, J. J. (2009). Out of Africa: modern human origins special feature: the origin of Neandertals. *Proceedings of the National Academy of Sciences of the United States of America, 106*(38), 16022-16027.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320.

Ionita-Laza, I., Rogers, A. J., Lange, C., Raby, B. A., Lee, C. (2009). Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics, 93*(1), 22-26.

Jeffreys, A. J., Pena, S. D. (1993). Brief introduction to human DNA fingerprinting. *EXS, 67*, 1-20.

Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L., Armour, J. A. (1994). Complex gene conversion events in germline mutation at human minisatellites. *Nature Genetics, 6*(2), 136-145.

Jeffreys, A., Wilson, V., Thein, S. (1985). Individual-Specific Fingerprints of Human Dna. *Nature, 316*(6023), 76-79.

Jurka, J. (1993). A new subfamily of recently retroposed human Alu repeats. *Nucleic Acids Research, 21*(9), 2252.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences of the United States of America, 94*(5), 1872-1877.

Jurka, J., Gentles, A. J. (2006). Origin and diversification of minisatellites derived from human Alu sequences. *Gene, 365*, 21-26.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research, 110*(1-4), 462-467.

Jurka, J., Klonowski, P., Trifonov, E. N. (1998). Mammalian retroposons integrate at kinkable DNA sites. *Journal of Biomolecular Structure & Dynamics, 15*(4), 717-721.

Jurka, J., Milosavljevic, A. (1991). Reconstruction and analysis of human Alu genes. *Journal of Molecular Evolution, 32*(2), 105-121.

Kajikawa, M., Okada, N. (2002). LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell, 111*(3), 433-444.

Kapitonov, V., Jurka, J. (1996). The age of Alu subfamilies. *Journal of Molecular Evolution, 42*(1), 59-65.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research, 12*(4), 656-664.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T. et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature, 453*(7191), 56-64.

Kim, U. K., Jorgenson, E., Coon, H., Leppert, M., Risch, N., Drayna, D. (2003). Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. *Science (New York, N.Y.), 299*(5610), 1221-1225.

Knebelmann, B., Forestier, L., Drouot, L., Quinones, S., Chuet, C., Benessy, F. et al. (1995). Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. *Human Molecular Genetics, 4*(4), 675-679.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F. et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science, 318*(5849), 420-426.

Krause, J., Orlando, L., Serre, D., Viola, B., Prufer, K., Richards, M. P. et al. (2007). Neanderthals in central Asia and Siberia. *Nature, 449*(7164), 902-904.

Kreahling, J., Graveley, B. R. (2004). The origins and implications of Aluternative splicing. *Trends in Genetics : TIG, 20*(1), 1-4.

Lam, H. Y., Mu, X. J., Stutz, A. M., Tanzer, A., Cayting, P. D., Snyder, M. et al. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology, 28*(1), 47-55.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. et al. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921.

Lee, S., Cheran, E., Brudno, M. (2008). A robust framework for detecting structural variations in a genome. *Bioinformatics (Oxford, England), 24*(13), i59-67.

Levin, H. L., Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. *Nature Genetics, 12*, 615-627.

Levinson, G., Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution, 4*(3), 203-221.

Lev-Maor, G., Sorek, R., Shomron, N., Ast, G. (2003). The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science (New York, N.Y.), 300*(5623), 1288-1291.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P. et al. (2007). The diploid genome sequence of an individual human. *PLoS Biology, 5*(10), e254.

Li, H., Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England), 25*(14), 1754-1760.

Lunt, D. H., Whipple, L. E., Hyman, B. C. (1998). Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. *Molecular Ecology, 7*(11), 1441-1455.

Luo, X., Dehne, F., Liang, P. (2011). Identification of Transposon Insertion Polymorphisms (TIPs) by computational comparative analysis of next generation personal genome data. *AIP Conference Proceedings, 1368*, 163-166.

Mandel, J. L. (1997). Human genetics. Breaking the rule of three. *Nature, 386*(6627), 767-769.

Martienssen, R. A. (2003). Maintenance of heterochromatin by RNA interference of tandem repeats. *Nature Genetics, 35*(3), 213-214.

Mashkova, T. D., Oparina, N. Y., Lacroix, M. H., Fedorova, L. I., G Tumeneva, I., Zinovieva, O. L., Kisselev, L. L. (2001). Structural rearrangements and insertions of dispersed elements in pericentromeric alpha satellites occur preferably at kinkable DNA sites. *Journal of Molecular Biology, 305*(1), 33-48.

Matera, A. G., Hellmann, U., Schmid, C. W. (1990). A transpositionally and transcriptionally competent Alu subfamily. *Molecular and Cellular Biology, 10*(10), 5424-5432.

Mathias, S. L., Scott, A. F., Kazazian, H. H.,Jr, Boeke, J. D., Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science (New York, N.Y.), 254*(5039), 1808-1810.

McClintock, B. (1956). Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology, 21*, 197-216.

McDougall, I., Brown, F. H., Fleagle, J. G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature, 433*(7027), 733-736.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F. et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research, 19*(9), 1527-1541.

Medvedev, P., Stanciu, M., Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods, 6*(11 Suppl), S13-20.

Mellars, P. (2006). Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proceedings of the National Academy of Sciences of the United States of America, 103*(25), 9381-9386.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews.Genetics, 11*(1), 31-46.

Mills, R. E., Bennett, E. A., Iskow, R. C., Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics : TIG, 23*(4), 183-191.

Mills, R. E., Bennett, E. A., Iskow, R. C., Luttig, C. T., Tsui, C., Pittard, W. S., Devine, S. E. (2006). Recently mobilized transposons in the human and chimpanzee genomes. *American Journal of Human Genetics, 78*(4), 671-679.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C. et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature, 470*(7332), 59-65.

Miyamoto, M. M., Slightom, J. L., Goodman, M. (1987). Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science (New York, N.Y.), 238*(4825), 369-373.

Mizuuchi, K. (1992). Transpositional recombination: mechanistic insights from studies of mu and other elements. *Annual Review of Biochemistry, 61*, 1011-1051.

Morris, C. A., Moazed, D. (2007). Centromere assembly and propagation. *Cell, 128*(4), 647-650.

Muratani, K., Hada, T., Yamamoto, Y., Kaneko, T., Shigeto, Y., Ohue, T. et al. (1991). Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. *Proceedings of the National Academy of Sciences of the United States of America, 88*(24), 11315-11319.

Murray, J., Buard, J., Neil, D. L., Yeramian, E., Tamaki, K., Hollies, C., Jeffreys, A. J. (1999). Comparative sequence analysis of human minisatellites showing meiotic repeat instability. *Genome Research, 9*(2), 130-136.

Nei, M., Roychoudhury, A. K. (1993). Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution, 10*(5), 927-943.

Nishizawa, S., Kubo, T., Mikami, T. (2000). Variable number of tandem repeat loci in the mitochondrial genomes of beets. *Current Genetics, 37*(1), 34-38.

Novick, G. E., Gonzalez, T., Garrison, J., Novick, C. C., Batzer, M. A., Deininger, P. L., Herrera, R. J. (1993). The use of polymorphic Alu insertions in human DNA fingerprinting. *EXS, 67*, 283-291.

Novick, G. E., Novick, C. C., Yunis, J., Yunis, E., Martinez, K., Duncan, G. G. et al. (1995). Polymorphic human specific Alu insertions as markers for human identification. *Electrophoresis, 16*(9), 1596-1601.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biology, 4*(11), R74.

Olaisen, B., Stenersen, M., Mevag, B. (1997). Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics, 15*(4), 402-405.

Oldridge, M., Zackai, E. H., McDonald-McGinn, D. M., Iseki, S., Morriss-Kay, G. M., Twigg, S. R. et al. (1999). De novo alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *American Journal of Human Genetics, 64*(2), 446-461.

Ostertag, E. M., Goodier, J. L., Zhang, Y., Kazazian, H. H.,Jr. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. *American Journal of Human Genetics, 73*(6), 1444-1451.

Ostertag, E., Kazazian, H. (2001). Biology of mammalian L1 retrotransposons. *Annual Review of Genetics, 35*, 501-538.

Pearce, E., Stringer, C., Dunbar, R. I. (2013). New insights into differences in brain organization between Neanderthals and anatomically modern humans. *Proceedings.Biological Sciences / the Royal Society, 280*(1758), 20130168.

Plohl, M., Mestrovic, N., Mravinac, B. (2012). Satellite DNA evolution. *Genome Dynamics, 7*, 126-152.

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S. et al. (2006). Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics, 2*(10), e168.

Prabhakar, S., Noonan, J. P., Paabo, S., Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science (New York, N.Y.), 314*(5800), 786.

Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A. et al. (2008). Human-specific gain of function in a developmental enhancer. *Science (New York, N.Y.), 321*(5894), 1346-1350.

Price, A. L., Eskin, E., Pevzner, P. A. (2004). Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Research, 14*(11), 2245-2252.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England), 28*(18), i333-i339.

Rebollo, R., Romanish, M. T., Mager, D. L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics, 46*, 21-42.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D. et al. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454.

Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M. et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science (New York, N.Y.), 316*(5822), 222-234.

Richard, G. F., Paques, F. (2000). Mini- and microsatellite expansions: the recombination connection. *EMBO Reports, 1*(2), 122-126.

Roy, A. M., Carroll, M. L., Kass, D. H., Nguyen, S. V., Salem, A. H., Batzer, M. A., Deininger, P. L. (1999). Recently integrated human Alu repeats: finding needles in the haystack. *Genetica, 107*(1-3), 149-161.

Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A. H., Oldridge, M., Wilkie, A. O. et al. (2000). Potential gene conversion and source genes for recently integrated Alu elements. *Genome Research, 10*(10), 1485-1495.

Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H. et al. (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics, 159*(1), 279-290.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G. et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature, 409*(6822), 928-933.

Saitou, N., Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406-425.

Schlotterer, C., Vogl, C., Tautz, D. (1997). Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural Drosophila melanogaster populations. *Genetics, 146*(1), 309-320.

Schmid, C. W. (1993). How many source Alus? *Trends in Genetics : TIG, 9*(2), 39.

Shen, L., Wu, L. C., Sanlioglu, S., Chen, R., Mendoza, A. R., Dangel, A. W. et al. (1994). Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *The Journal of Biological Chemistry, 269*(11), 8466-8476.

Shen, M. R., Batzer, M. A., Deininger, P. L. (1991). Evolution of the master Alu gene(s). *Journal of Molecular Evolution, 33*(4), 311-320.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research, 29*(1), 308-311.

Smit, A. F. A., Hubley, R. & Green, P. Retrieved from http://www.repeatmasker.org

Spurr, N. K., Bryant, S. P., Attwood, J., Nyberg, K., Cox, S. A., Mills, A. et al. (1994). European Gene Mapping Project (EUROGEM): genetic maps based on the CEPH reference families. *European Journal of Human Genetics : EJHG, 2*(3), 193-203.

Stam, M., Belele, C., Dorweiler, J. E., Chandler, V. L. (2002). Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. *Genes & Development, 16*(15), 1906-1918.

Stewart, C., Kural, D., Stromberg, M. P., Walker, J. A., Konkel, M. K., Stutz, A. M. et al. (2011). A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genetics, 7*(8), e1002236.

Stringer, C. B., Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science (New York, N.Y.), 239*(4845), 1263-1268.

Stringer, C. B., Hublin, J. (1999). New age estimates for the Swanscombe hominid, and their significance for human evolution. *Journal of Human Evolution, 37*(6), 873-877.

Sukarova, E., Dimovski, A. J., Tchacarova, P., Petkov, G. H., Efremov, G. D. (2001). An Alu insert as the cause of a severe form of hemophilia A. *Acta Haematologica, 106*(3), 126-129.

Sutherland, G., Baker, E., Richards, R. (1998). Fragile sites still breaking. *Trends in Genetics, 14*(12), 501-506.

Tamaki, K., Huang, X. L., Yamamoto, T., Uchihi, R., Nozawa, H., Katsumata, Y. (1995). Applications of minisatellite variant repeat (MVR) mapping for maternal identification from remains of an infant and placenta. *Journal of Forensic Sciences, 40*(4), 695-700.

Tamura, K., Nei, M., Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America, 101*(30), 11030-11035.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution, 28*(10), 2731-2739.

Taylor, J. S., Breden, F. (2000). Slipped-strand mispairing at noncontiguous repeats in Poecilia reticulata: a model for minisatellite birth. *Genetics, 155*(3), 1313-1320.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M. et al. (2005). Fine-scale structural variation of the human genome. *Nature Genetics, 37*(7), 727-732.

van de Lagemaat, L. N., Gagnier, L., Medstrand, P., Mager, D. L. (2005). Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Research, 15*(9), 1243-1249.

van Luenen, H. G., Colloms, S. D., Plasterk, R. H. (1994). The mechanism of transposition of Tc3 in C. elegans. *Cell, 79*(2), 293-301.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. et al. (2001). The sequence of the human genome. *Science (New York, N.Y.), 291*(5507), 1304-1351.

Vervoort, R., Gitzelmann, R., Lissens, W., Liebaers, I. (1998). A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene. *Human Genetics, 103*(6), 686-693.

Wada, C., Shionoya, S., Fujino, Y., Tokuhiro, H., Akahoshi, T., Uchida, T., Ohtani, H. (1994). Genomic instability of microsatellite repeats and its association with the evolution of chronic myelogenous leukemia. *Blood, 83*(12), 3449-3456.

Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., Walker, J. A., Batzer, M. A. (2005). SVA elements: a hominid-specific retroposon family. *Journal of Molecular Biology, 354*(4), 994-1007.

Wang, J., Song, L., Gonder, M. K., Azrak, S., Ray, D. A., Batzer, M. A. et al. (2006a). Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. *Gene, 365*, 11-20.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M. A., Liang, P. (2006b). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation, 27*(4), 323-329.

Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T., Toyoda, A., Kuroki, Y. et al. (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22 RID A-7121-2009 RID E-3997-2010. *Nature, 429*(6990), 382-388.

Weiner, A. M., Deininger, P. L., Efstratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annual Review of Biochemistry, 55*, 631-661.

Wetterbom, A., Sevov, M., Cavelier, L., Bergstrom, T. F. (2006). Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *Journal of Molecular Evolution, 63*(5), 682-690.

Xie, C., Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics, 10*, 80-2105-10-80.

Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C. H. et al. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell, 153*(4), 919-929.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research, 19*(9), 1586-1592.